DOCUMENT RESUME

ED 150 167 TM 006 860

AUTHOR	Fink, C. Dennis; And Others
TITLE	A State-of-the-Art Review of Techniques and Procedures for the Measurement of Complex Human Performance. Consulting Report.
INSTITUTION	Human Resources Research Organization, Alexandria, Va.
SPONS AGENCY	National Board of Medical Examiners, Philadelphia, Pa.
REPORT NO	HumRRO-CR-D1-74-2
PUB DATE	Apr 74
CONTRACT	574-11
NOTE	208p.
EDRS PRICE	MF-$0.83 HC-$11.37 Plus Postage.
DESCRIPTORS	Achievement Tests; Aptitude Tests; Behavioral Objectives; Bibliographies; Communication Skills; Decision Making Skills; *Evaluation Methods; Industry; *Informal Assessment; *Literature Reviews; Measurement Techniques; Medical Care Evaluation; Military Organizations; Perception; *Performance Tests; *Personnel Evaluation; Physicians; Psychomotor Skills; Rating Scales; Records (Forms); Reliability; Self Evaluation; *State of the Art Reviews; Testing Programs; Validity

ABSTRACT

Recent efforts to assess complex human performances in various work settings are reviewed. The review is based upon recent psychological, educational, and industrial literature, and technical reports sponsored by the military services. A few selected military and industrial locations were also visited in order to learn about current research and development efforts related to personnel performance evaluation. Although this survey was concerned primarily with techniques for assessing the performance of trained practitioners, as opposed to aptitude testing, certain aptitude measurement procedures which could be applied to the post-training assessment of behavior were also included. Measurement techniques which are typically employed to assess trainees but could also be used to assess practitioners are also described. Techniques for assessing perceptual processes, communication or interpersonal skills, and psychomotor skills are included. The evaluation methods are classified under the headings: achievement and aptitude tests (multiple choice tests), self-provided information (interviews and self-appraisals), appraisal by others (rating and ranking techniques), behavioral measurement (situational proficiency tests), and record audit (medical or flight records audit). For each method, its reliability, validity, current state, feasibility, and sources of additional information are discussed. Its appropriateness for evaluation of physicians is also considered. (Author/MV)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED150167

CR-D1-74-2

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Human Resources Research Organization

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

CONSULTING REPORT

A State-of-the-Art Review of Techniques and Procedures for the Measurement of Complex Human Performance

Prepared for

THE NATIONAL BOARD OF MEDICAL EXAMINERS
Philadelphia, Pennsylvania

Prepared by

HUMAN RESOURCES RESEARCH ORGANIZATION
Alexandria, Virginia

April 1974

C. Dennis Fink
Harold Wagner
Richard D. Behringer
Francis L. Hibbits

TM006 860

CR-D1-74-2

CONSULTING REPORT

A State-of-the-Art Review of Techniques and Procedures for the Measurement of Complex Human Performance

Prepared for

THE NATIONAL BOARD OF MEDICAL EXAMINERS
Philadelphia, Pennsylvania

Prepared by

HUMAN RESOURCES RESEARCH ORGANIZATION
Alexandria, Virginia

April 1974

C. Dennis Fink
Harold Wagner
Richard D. Behringer
Francis L. Hibbits

FOREWORD

The National Board of Medical Examiners currently is interested in how its testing programs can be expanded to assess more adequately non-cognitive processes. Prior to undertaking an intensive review and investigation of techniques for assessing non-cognitive process it was decided to conduct a state-of-the-art review of recent assessment practices. This report briefly reviews and presents judgmental comments pertaining to recently reported or currently employed techniques for assessing human performance.

This study was performed by Division No. 1 of the Human Resources Research Organization (HumRRO), Alexandria, Virginia, Dr. J. Daniel Lyons, Division Director. The study, sponsored by the National Board of Medical Examiners, commenced in October 1973 and continued into March 1974.

C. Dennis Fink, Ph.D. was the Principal Investigator. Dr. Fink was assisted by Harold Wagner, Ph.D., Richard D. Behringer, Ph.D., and Mr. Francis L. Hibbits all of HumRRO Division No. 1

This report was prepared for the internal use of the National Board of Medical Examiners, to include any of its Advisory Committees.

This report was prepared pursuant to HumRRO Contract No. 574-11 with the National Board of Medical Examiners, Philadelphia, Pa.

Meredith P. Crawford
President
Human Resources Research Organization

TABLE OF CONTENTS

	Page
INTRODUCTION	1
SURVEY FRAMEWORK AND PROCEDURES	2
Delimitation of Search Areas	2
Classification of Methods	3
Search Areas	7
Method Descriptions	8
DISCUSSION OF FINDINGS	13
Educational & Psychological vs. Job Performance Assessment	13
Student vs. Practitioner Control	14
Opinions About Assessment Program Costs	15
Large vs. Small-Scale Assessment Programs	16
Job Criteria	16
Visits to Military and Industrial Organizations	17
Use of This Report	17
ASSESSMENT METHODS AND APPROACHES	19
Measurement Category I: Self-Provided Information	19
Measurement Category II: Appraisal by Others	31
Measurement Category III: Behavioral Measurement	77
Measurement Category IV: Behavioral Record Audit	157
ADDITIONAL REFERENCES	161

TABLES

		Page
1	Classification of Behaviors	4
2	Measurement Categories	6
3	A Classification of Behavioral Measurement Techniques in Terms of Type of Process Assessed, Directness of Measurement, and Nature of Assessment Environment	9
4	A Classification of Behavioral Measurement Techniques in Terms of Their Current State of Refinement and Their Judged Suitability for Use with a Nationwide Testing Program	10

INTRODUCTION

This report describes the findings of "A State-of-the-Art Review of Techniques and Procedures for the Measurement of Complex Human Performance". It has been prepared for the internal use of the Committee on Research and Development of the National Board of Medical Examiners (NBME).

In 1971 the NBME established a Committee on Goals and Priorities. This Committee was charged with the mission of reviewing graduate medical education evaluation and "to develop a report 'suggesting goals and priorities of the NBME for a ten-year period'." One recommendation stemming from that report was that the NBME should establish an Office of Research and Development. Among other responsibilities this Office "would be responsible for the implementation of the overall research program (of the NBME) including the planning, development, coordination, and funding of all research activities."

To assist the NBME in planning its initial R&D efforts, it has commissioned the preparation of two state-of-the-art reports. The first report, authored by Paul J. Sanazaro, M.D. and John S. Lloyd, Ph.D. described the "results of a survey and analysis of selected current and past efforts to assess physician clinical performance." This report describes recent efforts to assess complex human performance in non-medical settings. The report is based on a review of recent psychological and educational literature, industrial literature, and technical reports and regulations sponsored or prepared by the Military Services.

A few selected military installations and industrial concerns also were visited to learn about current R&D efforts related to human performance assessment.

SURVEY FRAMEWORK AND PROCEDURES

Delimitation of Search Area

One or more techniques have been developed for the measurement of each of a countless number of specific or general behaviors, attitudes, aptitudes and traits. To restrict the scope of this study it was decided that methods concerning the measurement of Mediational or Cognitive processes would not be surveyed. These processes are assessed quite well by current NBME testing procedures and devices. Also, techniques concerned with the measurement of attitudes and traits were not reviewed.

This survey was concerned primarily with techniques for assessing the performance of trained persons and practitioners as opposed to techniques for assessing performance aptitudes. However, certain techniques, developed primarily to assess aptitudes, also can or might be used to assess the performance of practitioners or of advanced trainees. Therefore, this report will describe certain aptitude assessment procedures which, in our judgment, could be applied to the post-training assessment of behavior.

A number of the techniques reported on typically are employed to assess trainees, and depend for their success on the simulation of certain features of a work environment. These techniques also could be used to assess practitioners.

The areas surveyed can best be described with reference to a taxonomic scheme developed by Berliner and his associates. This scheme (see Table 1) classifies behavior into four major processes encompassing six broad types of activities under which are subsumed some 47 specific behaviors. This study surveyed techniques for assessing perceptual processes, communication (interpersonal) processes, and motor (psychomotor) processes.

Classification of Methods

Behavioral measurement methods and techniques can be classified in a variety of ways. The classification scheme used in this study was adapted from one employed by Thorndike and Hagen[1]. According to this scheme, measurement methods can be classified under the following five headings:

a. Achievement and Aptitude Tests (eg., multiple-choice tests)

b. Self-provided information (eg., interviews & self-appraisals)

c. Appraisal by others (eg., rating and ranking techniques)

d. Behavioral Measurement (eg., situational proficiency tests)

e. Record Audit (eg., medical records audit & flight recorder audit)

With a few exceptions paper and pencil test-related methods will not be reported on.

---

[1] Thorndike, Robert L. and Hagen, Elizabeth. Measurement and Evaluation in Psychology and Education. (3rd Ed.) John Wiley & Son, Inc. New York, 1969.

TABLE I

Classification of Behaviors[1]

PROCESSES	ACTIVITIES	SPECIFIC BEHAVIORS
Perceptual Processes	Searching for and Receiving Information	Detects, Inspects, Observes, Reads, Receives, Scans, Surveys
	Identifying Objects, Actions, Events	Discriminates, Identifies, Locates
Mediational Process	Information Processing	Categorizes, Calculates, Codes, Computes, Interpolates, Itemizes, Tabulates, Translates
	Problem Solving and Decision Making	Analyzes, Calculates, Chooses, Compares, Computes, Estimates, Plans
Communication Processes		Advises, Answers, Communicates, Directs, Indirects, Informs, Instructs, Requests, Transmits
Motor Processes	Complex/Continuous	Adjusts, Aligns, Regulates, Synchronizes, Tracks
	Simple/Discrete	Activates, Closes, Connects, Disconnects, Joins, Moves, Presses, Sets

[1] Taken from Angell, David, Shearer, J.W. and Berliner, D.C., Study of Training Performance Evaluation Techniques, Technical Report NAVTRADECEN 1449-1, 74 pg. 16 October 1964.

Within each major methods category there falls a variety of techniques many of which can be grouped into distinct sub-categories. To further complicate matters, some sub-categories of techniques can be divided into sub-sub-categories. For example, we have sub-divided the "Behavioral Measurement" category into four sub-categories, one of which is labeled "Systematic Observation." This sub-category contains over ninety different assessment techniques, many of which can be grouped into sub-sub-categories. Table 2 shows the arrangement of categories we have developed for classifying the techniques surveyed during this study. It should be recognized that many techniques are "mixed" in that they are not readily subsumed under a single rubric.

A number of other important dimensions could have been used to classify the techniques surveyed during this study. Some of these are:

a. direct vs. indirectness of assessment technique (is behavior directly or indirectly observed and/or measured)

b. process vs. outcome (does the technique assess how one arrives at an outcome, the nature and/or "goodness" of the outcome, or both)

c. natural vs. simulated testing environment (does the technique assess behavior in a natural work setting or in a simulated work environment)

b. same vs. cross-cultural setting (can the technique be used to assess behavior in a cross-cultural setting, or is it confined to a same-cultural setting)

When discussing individual assessment methods their relation to the above dimensions will be noted.

Table 2

	MEASUREMENT CATEGORIES	Index No.
I.	SELF-PROVIDED INFORMATION	
	A. Self-Ratings	1
	B. Confidence Testing	2
II.	APPRAISAL BY OTHERS	
	A. Rating Scales	
	1. Behavioral (Expectancy) Scales (BES)	3
	2. Forced-Choice Rating Scales	4
	3. Mixed Standard Scales	5
	B. Personnel Evaluation Systems and Approaches	
	1. Performance Evaluation System	6
	2. Military Officer Efficiency Reports	7
	3. Enlisted Performance Evaluation System	8
	4. Competency-Based Education and Certification Procedures	9
	5. The Multitrait-Multimethod (Rater) Approach	10
III.	BEHAVIORAL MEASUREMENT	
	A. Behavioral Tests	11
	1. Listening	12
	B. Situational Tests and Assessment Programs	
	1. The Assessment Center	13
	2. Leadership Assessment Simulations	14
	3. Military Human Performance Assessment Programs	15
	C. Systematic Observation Systems	
	1. Sequential Analysis of Verbal Interaction	16
	2. Flanders System of Interaction Analysis	17
	3. CERLI Verbal Classification System	18
	4. Medical Instruction Observation Record	19
	5. Notation of Proxemic Behavior	20
	6. Audio/Visual Record System for Pilot Training	21
	7. Interviewing	22
	D. Performance Testing Techniques	23
	1. Simulation Techniques	24
	a. Gaming	25
	b. Role-Playing	26
	c. In-Basket Exercise	27
	2. Natural Setting Techniques	28
	3. Adaptive Performance Testing	29
IV.	BEHAVIORAL RECORD AUDIT	
	Overview: Record Audit	30
	Personal Histories and Diaries	31
	Informal Observations & Anecdotal Records	32
	Unobtrusive Measures	33

Search Areas

This state-of-the-art survey was performed by first searching a variety of literature sources. This search was confined to literature reported from 1969 to the present. Following this, a few military and industrial organizations were visited in order to identify current R&D efforts in the area of human behavior assessment. The literature search encompassed:

a. A review of issues of Psychological Abstracts and ERIC Abstract Reports for the years 1969 through 1973.

b. A review of a number of bibliographies prepared for HumRRO during 1973. These were prepared by the National Technical Information Service and covered such topics as performance testing, certification, training program evaluation, and so on.

c. A review of training and evaluation-related technical reports produced by or for major military R&D organizations during the past five years. HumRRO is on the distribution list for most of these reports.

d. A review of current military regulations which describe the performance assessment procedures to be followed by each of the Services.

e. A review of standard trade and industrial journals, to include professional journals such as the Harvard Business Review, for articles related to performance assessment as practiced in business and industry.

The literature search resulted in the identification of more than 500 potentially useful journal articles and reports. Abstracts were obtained for these. Then, following further examination of the abstracts,

a sub-set of articles were identified which seemed to contain the best information about the techniques to be discussed in the report. Microfiche or hard-copy was obtained for each of these reports. These were studied and served to provide the information for preparing the Method Descriptions contained in this report.

Method Descriptions

For each of the method categories and sub-categories shown in Table 2, a Method Description has been prepared. These are located by Category after the report discussion section. The matrices shown in Tables 3 and 4, have been used to summarize certain characteristics of plus our judgments about each method. Other characteristics have been summarized at the end of each Method Description.

It should be noted that a Method Description may describe either a very general or a very specific assessment procedure. For example, many industrial firms have developed a set of procedures for evaluating their personnel. This class of procedures, commonly known as Performance Evaluation, is often discussed as though it was a single method. In fact, it is a conglomerate of methods tied together by a standardized set of administrative procedures. Thus, in this report a Method Description has been provided for Performance Evaluation as a unitary set of procedures. In addition, descriptions have been provided for many of the individual measurement techniques which comprise the more general method.

A Classification of Behavioral Measurement Techniques in Terms of Type of Process Assessed, Directness of Measurement, & Nature of Assessment Environment

Behavioral Process		Direct Measurement			Indirect Measurement		
		Actual Environment	Simulated Environment	Both	Actual Environment	Simulated Environment	Both
Perceptual	Post-Training	28	22,24,26	11,12,14,23, 29	1,3,4,5,6,7, 8,31,32,33	13,14	
Mediational	Post-Training				3,4,5,6,7,8, 29,30,31,32, 33	2,13,14,24, 25,27	23,29
Communication	Post-Training	16,17,18,19, 20,22,23,28	24 25,26	14,23	3,4,5,6,7,8, 29,30,31,32, 33	13,14	
Motor	Post-Training	28	21,24,26	11,14 23, 29	3,6,7,8,21 32,33		
Perceptual	During Training	28	24,26	9,11,12,15, 23,29			10
Mediational	During Training		24,26	9,15	28	2,24,25,27	23,29
Communication	During Training	16,17,18,19, 20,23,27,28	24,26	9,15,23			
Motor	During Training	28	21,24,26	9,11,15,23, 29			

[1]The numbers in the matrix cells refer to the "Index Numbers" presented in Table 2.

Table 3

A Classification of Behavioral Measurement Techniques in Terms of Their Current State of Refinement and Their Judged Suitability for Use with a Nationwide Testing Program.

Current State of Refineness		Suitability for Use in a Nationwide Physician Assessment Program			
		High	Moderately High	Moderately Low	Low
Very unrefined	Post-Training				7,31,32
Somewhat unrefined	Post-Training	30	5,12,22	6,8,23,29,33	28
Somewhat refined	Post-Training	2	4,21, 26	11,13,16,17,18, 19,20,23,24	1,9,14,15
Very refined	Post-Training		3,27	25	10
Very unrefined	During Training				
Somewhat unrefined	During Training	12	22,29	23	28
Somewhat refined	During Training	2	11,21,23,24,26	16,17,18,19,20	1,14
Very refined	During Training	25,27			

Table 4

Each Method Description contains a general description of the method, what types of behavior it can assess, the assessment environment, and so on. This is followed by a series of judgmental statements about the method. These statements are based on information obtained from the literature and on what we, the Contractor personnel, know or believe relative to the method. These judgmental statements pertain to the following areas:

a. Reliability of the Method

b. Validity of the Method

c. Current State of Refineness

d. Feasibility of Using the Method. A general estimate of:

   1) Costs
   2) People Requirements
   3) Test Environment Requirements
   4) Suitability for a Nation-Wide Testing Program

e. Type of Behavioral Process Typically Assessed by the Method

f. Appropriateness of NBME R&D Support. A general estimate of:

   1) Refinement Potential
   2) Adaptability to Assessment of Physicians
   3) Potential for Assessing Non-Cognitive Processes

Each Method Description terminates with a listing of information sources about the method, to include:

a. Recent Bibliographic References

b. Names and addresses of persons or organizations (other than reference authors) who might be contacted regarding the method.

The reviewers encountered certain assessment methods which are of doubtful utility to the NBME. Some of these methods are so unrefined that we would not wish to foster their use in a medical setting, i.e. Officer Efficiency Reports as employed by the Military Services. Other methods or versions thereof have been reviewed in recent textbooks, but since then little significant research has been reported on their use. The use of peer ratings seems to be one such example. Still other methods, such as the video recording of study pilot behavior, are recording methods employed with a wide assortment of assessment procedures. Methods following into any of these categories have been briefly summarized, but detailed judgmental statements about them have not been prepared.

DISCUSSION OF FINDINGS

"Applied psychologists, educators and training specialists have produced great numbers of reports on training assessment, but the vast majority use school graduation or classroom measuring instruments. To all intents and purposes one finds only isolated attempts to measure performance on the job by other than rating methods when that performance was intended to serve as a measure of the effectiveness of the training that preceded it." The foregoing, quoted from Clark L. Wilson's Chapter on "On-the-job and Operational Criteria," described the state of affairs in job performance measurement as of 1962. In the last eleven years the state-of-the-art has advanced, but very slowly. Regarding educational measurement, Thorndike reported in 1969 that "Developments in testing have been evolutionary rather than revolutionary... The art (of educational measurement) is growing, but it is the same basic art."

Educational & Psychological vs. Job Performance Assessment. By tradition most measurement and research has been related to education -- the selection and classification of students, and the assessment of student performance during or at the end of a course of instruction. Another group of measurement experts has been concerned with the development of psychological tests. Industrial and personnel psychologists make heavy use of selection, classification and personality measurement devices. However, they have not been able to develop similar devices to assess job performance. In part this is because job performance cannot easily be assessed by paper and pencil tests. Also, few industrial companies are willing to expend the time and money required to

develop and administer a wide variety of performance-oriented tests. The military services are willing to do this but for very special reasons which will be discussed in this report.

It would seem that educators should be interested in whether or not a well trained student also is one who can perform well on the job. Indeed they are interested in this relationship and try to ensure the relation by developing courses which teach job-related skills and knowledges. Educators and trainers may go to considerable length to assure that the training objectives of a course are related to the more important features of a class of job positions. After developing a course based on behaviorally stated training objectives one then can develop assessment procedures to determine if students can perform in accordance with training objective standards. Testing may occur in a simulated work environment and the test items may require the students to perform specific job-related tasks. Assuming these conditions to hold it can be argued that students who can perform well in the test environment also can perform well in the "real world."

If you cannot satisfactorily assess on-the-job performance then develop a training system which assures that students will be taught to perform adequately those tasks which are known (or thought) to be related to good job performance. This is the philosophy upon which competency-based instruction and certification systems are based. Such systems have been reviewed in this report.

Student vs. Practitioner Control. During training students are to some degree under the control of a school administration. Therefore they can be subjected to a variety of testing situations some of which

they might not tolerate once they became certified practitioners. The military services, however, have fairly tight control over both students and practitioners. Not surprisingly therefore, the Services are actively involved in the assessment of post-training proficiency.

For the most part military personnel perform within a unit. Therefore, post-training proficiency assessment often is in the form of unit proficiency assessment. For example, the Air Force might evaluate a flight crew by the use of a team of trained observer-raters who go along on a training flight. The Army might transport an entire missile firing team to a test site and require the team to put into order and fire missiles from a Nike-Hercules missile installation.

One of the main goals of the military establishment is to maintain itself in a state of readiness. Therefore, throughout their career military officers, and to a lesser extent enlisted personnel, are required to attend training programs. From time to time they also engage in unit training exercises. The most sophisticated assessment procedures developed by the military are associated with these training programs and exercises. In one sense military personnel remain in a student status until it becomes time for them to practice the art of fighting. At that time assessment of their job performance capability ceases.

Opinions about Assessment Program Costs. It seems to be commonly accepted that considerable training equipment and simulation facilities may be needed to conduct training for certain types of tasks. However, decision makers, especially accountants, are much less convinced that considerable funds should be spent to assess the performance of those

who apparently already are successful practitioners. One outgrowth of these opinions has been the attempt to assess practitioner performance by the use of elaborate simulations and simulators developed for use during training. These attempts may not be too successful. Simulations developed for training purposes often are not wholely appropriate for the assessment of practitioners. This is because training simulators usually allow only for the execution of a limited range of job-related responses to a limited range of job-related stimuli.

Large vs. Small-Scale Assessment Programs. Both industry and the Military Services have been using job performance assessment systems for many years. These systems usually are based on the use of poorly conceived rating scales. It seems that cost and feasibility considerations have prevented the development of better programs. As a result assessment program refineness has been degraded to a point where program costs and administration efforts became tolerable.

Job Criteria. It is not too difficult for a group of knowledgeable experts to make reasonable guesses regarding the traits which should be possessed by a good job performer. Hopefully, statistical techniques then can be used to identify the few factors which "really are" related to job performance. In a recent text on educational measurement (Thorndike, 1971), a number of promising statistical developments are discussed in quite some detail. In the meanwhile others are attempting to address the criterion problem more directly using groups of content experts to define dimensions of job-related behavior. These techniques are used to develop Behavioral Expectancy Scales and Forced-Choice

Scales, both reviewed in this report. The problem of identifying criterion behavior is discussed more fully in the Overview to Rating Scales.

Visits to Military and Industrial Organizations. The visitations made during the preparation of this report were less than originally anticipated. A review of military service assessment approaches led us to realize that Service assessment procedures are closely interwoven with a continuing training system and a system of announced and unannounced drills, alerts and unit training exercises. Our discussions with military personnel and our perusal of military-sponsored technical reports has led us to conclude that what is now going on with respect to human assessment in the Services is but an extension of what has been reported on in the last few years.

Assessment procedures employed by industrial concerns are not especially appealing to the purist. Industry is concerned with the cost-effectiveness of any approach; with respect to human performance assessment they seem to be mostly concerned with costs. The refinement of the "assessment center" approach by AT&T and others is one of the more reported-on assessment-related recent events in business and industry. However, a more substantive accomplishment would seem to be the development of better rating scales such as the Behavioral (Expectancy) Scales. Industry does employ a wide variety of assessment procedures. Most of these have been reported on during recent years.

Use of This Report. This report was prepared to assist members of the NBME as they begin the process of identifying those assessment techniques and approaches to examine in detail. To begin the detailed

study of any of the methods or approaches discussed herein, the following steps are suggested:

1. Read about the method in a recent text such as Measurement and Evaluation in Psychology and Education (Thorndike & Hagen), or Managerial Behavior, Performance and Effectiveness (Campbell, et. al).
2. Study one or more of the references listed after the description of the method or approach under study.
3. Read one or more of the additional references listed at the end of the report.
4. Contact one or more authors of recent relevant reports and arrange to visit them. With rare exceptions, the persons listed as report authors are the ones who should be contacted for additional information.

A number of Technical Reports have been listed as references. Much of the research sponsored by the Military Services is reported in such reports. These reports are in the public domain and may be purchased from the National Technical Information Center in the form of hard-copy or microfiche. Most of the cited technical reports are contained in the HumRRO library and can be obtained through interlibrary loan. HumRRO is ready to assist the NBME in obtaining any government-sponsored report which has been cleared for unlimited distribution.

ASSESSMENT METHODS AND APPROACHES

Measurement Category I: SELF-PROVIDED INFORMATION

Contents

	Index No.	Page No.
Overview		20
Self-Ratings	1	21
Confidence Testing	2	24

Overview

Self-provided information as discussed in this report includes all methods and techniques that an individual may use for self-appraisal or self-evaluation. Questionnaires, Rating Scales, Interviews and anecdotal information such as diaries are some of the means of supplying self-provided information.

Individuals can make self-evaluations about any type of behavior that they have performed or are expected to perform. Additionally, after actually performing the behavior they may express feelings of confidence concerning their own performance. This aspect of self-provided information will be discussed later in this report under the heading of "Confidence Testing."

Self-ratings of actual performance are commonly obtained in lieu of objective appraisal measures. Also, self-ratings may be used for self-improvement in training situations. Although used extensively and successfully for training purposes, it is unlikely that self-provided information can be used for assessment of behavior in a test situation. Self-evaluation simply cannot be substituted for objective assessment.

It would seem that physicians should be able to evaluate their performance realistically. The importance attached to this, however, relates to self-improvement, not to objective appraisal of performance.

Description

Self-ratings includes all methods that have been used for self-evaluative purposes. This category also may include level of aspiration measures or other self-predictions of performance.

Self-ratings of performance are often used when appraisal of performance by others is not feasible. The method also may be used after feedback of an individual's performance is provided for self-evaluation purposes under training conditions.

Microteaching is very popular and useful technique which involves the use of self-ratings. "Microteaching is a training concept that can be applied at various pre-service and in-service stages in the professional development of teachers. Microteaching provides a practice setting for instructing in which the normal complexities of the classroom are reduced and in which the teacher receives a great deal of feedback on his performance. To minimize the complexities of the normal teacher encounter, several dimensions are limited. The length of the lesson is reduced. The scope of the lesson is narrowed. In microteaching the teacher instructs only a few students instead of the normal 25 or 30 (Allen, et al, 1969)." Thus, the microteaching setting is a natural although miniaturized one and allows for self-ratings by the teacher for purposes of improving performance.

Self-ratings have been used to predict performance when the items rated are of low visibility -- when they have little face validity. For example, Robinson (1970) predicted police effectiveness from self-reports of relative time spent in task performance.

In at least one other study (Egelston, 1972) it has been found that students learn to improve their evaluation of self-performance on cognitive tasks. High-achieving students are more accurate at self-evaluation and improve at a faster rate than low-achieving students.

Among the many techniques used for self-ratings are rating scales, questionnaires, interview situations as well as other measures of self-report. Self evaluation has been accomplished under live conditions and using audio-visual tape playbacks.

Judgments About Technique

1. Reliability of Method. Unknown.
2. Validity of Method. Usually content valid.
3. Current State of Refineness. Fairly refined.
4. Feasibility of Using Method
   a. Cost. Minimal; only requires time for administration of self-rating scale or questionnaires.
   b. People Requirement. One person can administer instrument to several respondents.
   c. Test Environment Requirement. Must provide situation for individual to perform before self-evaluation can occur.
   d. Suitability for Nationwide Testing Program. Easy to handle administratively.
5. Type of Behavioral Process Typically Assessed by Method. Any kind of behavior is subject to self-evaluation.

6. Appropriateness of NBME R&D Support

a. Refinemenet Potential. Good.

b. Adaptability to Assessment of Physicians. Not applicable for objective assessment.

c. Potential for Assessing Non-Cognitive Processes. Good. Self-assessment of any behavior can be accomplished.

References

1. Allen, Dwight, et al. Microteaching, Addison-Wesley Publishing Co., Inc. Pg 151, Reading, Mass. 1969.

2. Robinson, David D. Predicting Police Effectiveness from Self Reports of Relative Time Spent in Task Performance. Personnel Psychology, 1970, Vol. 23, No. 3.

3. Egelston, Richard L., and Egelston, Judy C. Self-Evaluation and Performance on Classroom Tests, 1972, Paper presented at the National Association for Research in Science Teaching meeting, Chicago, Ill.

Description

Confidence testing refers to methods of weighting responses so as to reflect the examinee's belief in the correctness of his responses. The intent of these methods is to maximize the amount of information that can be obtained from a given set of test items. These methods differ primarily in the use of a variety of scoring systems by which the responses to test items are weighted and/or combined. Confidence testing has been used in varying forms over the past 40 years. In almost all cases, it has been used for tests of cognitive or mediational abilities. It is being discussed in this report because of its apparent applicability in any testing situation requiring discrete responses.

Scoring Methods

There are basically two different scoring methods that have been applied in confidence testing. One is called the "pick-one" method. Using this method, the student is asked to pick the answer he thinks is correct and then mark how confident he is that the answer he has chosen is, in fact, correct. The degree of confidence can be described in words (e.g., not sure, sure, very sure, etc.) or in the assignment of numbers which reflect the verbal scale. Scoring tables are then set up and developed so that the student will score highest if he gives a "true" estimate of his degree of certainty in his responses.

In another scoring system called, "distribute 100 points" the examinee indicates his confidence in each of the alternatives presented

him on the test question. Variations of this method employ fewer points such as 5 or 10, or the use of a mechanical device that helps the examinee allocate his confidence (SCoRule).*

We favor the "pick-one" scoring system because it requires a much easier task of the examinee than the other approach and more closely resembles testing situations which the examinee has experienced in the past. The scoring is not too difficult and the examinee is more likely to understand the outcome of his confidence rating.

The only extensive use of confidence testing reported in the literature has been on an experimental basis in technical training programs for the Air Force. Proponents of the use of confidence testing have cited the following advantages for its continuation:

1. Confidence testing may provide more information about an examinee's state of knowledge than does standard multiple choice or other paper and pencil tests.
2. The instructor can use the information obtained from confidence testing to prescribe instruction tailored to the needs of each individual student.
3. The use of confidence testing requires each examinee to be more careful before responding to each question. This supposedly increases their sensitivity to the content of the questions.

---

*SCoRule is a product of the Shuford-Massengill Corporation.

Judgments About Technique

1. Reliability. Results have shown a higher degree of reliability from confidence testing approaches than from multiple choice testing for the same number of items. In one example, reliability was increased from .79 to .85. Different scoring techniques have correlated approximately .9 on the same tests indicating that the choice of scoring method need not be based upon reliability.
2. Validity. Unknown. However, scoring systems have been developed which weight an individual's responses so that it "pays" him to respond as honestly as he can regarding his degree of certainty in his answers. Thus, the term "valid confidence testing" is sometimes employed to refer to the construct validity of these measures.
3. Current State of Refineness. The technique, although experimental, is fairly well refined. Increased use by the Air Force should lead to further refinement of the scoring techniques.
4. Feasibility of Use.
   a. Cost. Unknown
   b. People Requirements. No additional personnel required above those necessary for traditional test administration.
   c. Test Environment Requirements. Additional materials would be required for assisting the examinee score his confidence. Also, the examinees must know and understand the

scoring system to be used. This requires special training time to teach the confidence scoring technique. It has been reported that about 20% more time is required to administer, and twice as much time to score confidence tests than the traditional paper and pencil tests. The amount of extra time depends upon which scoring system is employed.

d. Suitability for Nationwide Testing Program. Confidence testing is highly suitable for application to a nationwide testing program as it can be applied as an adjunct to current testing procedures.

e. Number of People Tested Each Year. Unknown.

5. Type of Behavioral Process Typically Assessed by Method. Commonly used to assess cognitive processes. However, it may be applied to tests of communication or perceptual processes in which the individual has an opportunity to voice or record his degree of confidence in the correctness of his answers.

6. Appropriateness of NBME R&D Support.

a. Refinement Potential. Considerable potential for developing a confidence test scoring system which can be easily learned by examinees and readily employed in testing situations.

b. Adaptability to MD Assessment. At present, can be readily adapted to the assessment of cognitive processes. This adaptation could most easily occur in the medical schools. The students could learn the scoring system, and obtain

experience using it in their course examinations. With a lack of experience on the part of the examinees, studies have shown that personality factors tend to contaminate confidence test results.

c. Potential for Assessing Non-Cognitive Processes.

Moderate. This technique could be adapted to the assessment of inter-personal communication processes or perceptual processes. The only requirement is that the examinee have an opportunity to express the degree of confidence he has in the answers he gives. It is unlikely that confidence testing could be used in the assessment of motor skills.

References:

1. Boldt, Robert A., A Simple Confidence Testing Format, AFHRL-TR-31, July, 1971.

2. Echternacht, Gary J., Sellman, Wayne S., Cpt., Boldt, Robert F., and Young, Joseph D., Cpt. USAF, An Evaluation of the Feasibility of Confidence Testing as a Diagnostic Aid in Technical Training, Educational Testing Service, Princeton, New Jersey, October, 1971, RB-71-51.

3. Rippey, Robert M., Rationale For Confidence-Scored Multiple-Choice Tests, Psychological Reports, 1970, Vol, 27, 91-98.

4. Samet, Michael G., Admissible Probability Measurement in Psychological and Educational Research, Review Paper, Department of Psychology, Tufts University, March, 1969.

5. Shuford, Emir H., Jr., Confidence Testing: A New Tool for Measurement, The Shuford-Massengill Corp., Lexington, Mass. 02173.

2. People and Places to Visit

It is our opinion that an Air Force Technical Training Center which employs confidence testing should be visited to see it in operation (e.g., Chanute Air Force Base - if it is still being used there).

Measurement Category II: APPRAISAL BY OTHERS

Contents

	Index No.	Page No.
RATING SCALES		32
Overview		32
Behavioral (Expectancy) Scales (BES)	3	39
Forced-Choice Rating Scales	4	43
Mixed Standard Scales	5	46
References: Rating Scales		49
PERSONNEL EVALUATION SYSTEMS AND APPROACHES		51
Overview		51
Performance Evaluation Systems	6	54
Military Officer Efficiency Reports	7	59
Enlisted Performance Evaluation Systems	8	62
Competency-Based Education and Certification Procedures	9	68
The Multitrait-Multimethod (Rater) Approach	10	72

Overview

Rating scales probably are the most frequently employed measurement device to assess human performance. Such scales form the core of most personnel assessment approaches used today in industry and the military services. The reason for this seems obvious -- rating scales are deceptively easy to develop and administer and usually have considerable face validity. It is now quite widely recognized that valid rating scales are extremely difficult or at least time-consuming to develop and that their use is fraught with all sorts of difficulties.

A brief review of a few of the problems and issues related to rating scale development and use seems appropriate. Rating scales may be designed to measure a specific dimension of behavior, e.g., leadership ability or a more global aspect of behavior such as "performance as a manager." In the absence of specific definitions each user of a rating scale must supply his own definition of the dimension rated by the scale and of what constitutes good and poor performance along that dimension. Usually scale points are defined with reference to some vague behavioral referent -- very poor, somewhat below average, average, somewhat above average, very good. In recent years increased attention has been given to: (a) identifying those important dimensions of behavior which should be assessed; (b) defining with specific behavioral examples various points on the rating scale; and (c) using knowledgeable practitioners rather than scale developers to provide scale point definitions. This attention seems to be the result of a reluctant admission that the

developers of assessment procedures must more directly address the problem of what constitutes criterion behavior -- for any particular job position what are the indices of good, poor and indifferent performance.

From the standpoint of administrative ease one would like a performance assessment system which: (a) assessed as few performance dimensions as possible; and (b) employed only a few assessment devices. This desire has led to attempts, via factor analysis, to identify those few dimensions of behavior which collectively can account for most, if not all, performance related to a specific job. Although based on acceptable logic this approach seems to lead to unacceptable results -- management behavior is factored into such a small number of dimensions that it becomes difficult if not impossible to meaningfully define the dimensions. This has led some persons to wonder if global ratings of performance might not be as valid as the sum of ratings obtained by assessing many different behavioral dimensions each related to the same job. Many measurement practitioners feel uncomfortable using global ratings, but realize that in practice the number of dimensions assessed and devices employed must be kept to a minimum.

Before using a rating scale a person should receive some training on its use. At the very least the user should be made aware of the behavioral representatives of each scale point. The fact that the scale developer and users of the scale may have different notions regarding the behavioral definition of scale points can be troublesome. It increases the necessity for training users. In recent years scale developers have tried to minimize this problem by asking those who eventually will use the scale to help develop it. These users are

asked to define the behavioral dimension to be rated and to provide behavioral examples which can be used to define the scale points.

For the most part a rater assesses a ratee on the basis of recent past performance. This means that the rater should have had an ample opportunity to observe those persons to be rated. This condition sometimes is difficult to meet. Some persons perform in an isolated environment (fighter aircraft pilots); some perform isolated from supervisors or peers (private physicians and teachers); it's seldom that all aspects of a person's performance can be observed by the same supervisor or peer. Schemes have been developed so that those who eventually must rate an employee keep a daily or weekly "critical incident" diary of that employee's behavior. This serves as a memory jogger at rating time. Many management rating systems assess product output as well as behavioral performance which tends to be a process assessment.

Ratings tend to be more reliable, and probably more valid also, when they can be obtained from more than one rater. Sometimes this can be accomplished by obtaining ratings from different categories of people -- supervisors, peers, and subordinates. These would all be ratings obtained from external observers. Self-ratings also can be obtained, but these often do not correlate highly with those obtained from external observers. Supervisors, peers, and subordinates each see a person perform from a different perspective and often are exposed to different aspects of that person's performance. It would seem , therefore, that a combination of supervisor-peer-subordinate ratings would provide the most valid (but not necessarily the most reliable) assessment of performance. Research findings seem to support this hypothesis

provided that all raters have been taught to rate in terms of the same behavioral definitions. A few recent studies have compared ratings obtained from professional raters and/or content experts with such persons as clients, students, and patients -- persons who have a subordinate position to the person being rated. In the absence of special training these subordinates seem to overreact -- their ratings are too lenient or too severe.

A rating study often yields scores which appear to be too high. Raters tend to be lenient when rating others thus artifically inflating the rating scores (the "leniency effect"). The "halo effect" is another well-known deficiency of rating scales -- the tendency to rate all dimensions of behavior on the basis of exceptionally good or poor performance on one or two dimensions. To counteract these effects a "hidden" rating scale can be developed. That is, the scale can be developed such that socially acceptable or exceptional performance rating responses or options are not easily identified. Thus, the scale ratings cannot easily be manipulated by the rater. A "forced-choice" tetrad rating scale is a type of "hidden" scale. A Behavioral Expectancy Scale" is an "open" scale. Its scale points are arranged in an ascending order with each point carefully defined by one or more examples of a specific type of behavior.

In some instances an assessment checklist may be more useful than a rating scale. In fact one can develop rating scales which, for administrative purposes, can be converted into a checklist. For scoring purposes the checklist then can be converted back to a rating scale.

A checklist might contain a list of short behavioral descriptions; the rater would be asked to indicate which of the listed activities the ratee recently engaged in. For each activity the rater might also indicate how often the ratee seemed to engage in that type of activity -- never, seldom, and so on. The advantage of a checklist is that the rater is not asked to judge the ratee; rather the rater reports on what the ratee did and did not do.

Checklists are used to assess on-going behavior. For example, the Military Services use student pilot performance sheets to assess the performance of student pilots. For each flight maneuver a sheet has been prepared containing a list of the things the student pilot should or might do. On some sheets there are places to record certain facts such as airspeed, compass direction, and so on. For many flight maneuvers the criterion behavior has been defined quite precisely. Thus, it is possible to develop detailed checklists which an observor-pilot can use to record what a student did during a flight check. This behavioral record can then be scored by comparing it against the criterion behavior. Before the development of these flight criteria and associated checklists the quality of performance for new pilots varied quite widely.

During the last five years not a great deal of significant work on the development and use of rating scales has been reported. A few studies have reported on the development and application of Behavioral Expectancy Scales or Forced-Choice scales. A few articles have discussed the advantages of using a multitrait-multirater approach to analyze assessment methods and data. A few other articles have promoted the use of "multidimensional scales" and "mixed" scales. These terms refer to the practice

of employing more than one method or measuring device to simultaneously assess more than one dimension of behavior. This practice is not much different from that which will be discussed under the label of "Assessment Centers" and "Performance Evaluation" systems. Cited briefly below are a few recent research findings which relate to the potential usefulness of rating scales to the assessment of medical practitioners. This will be followed by write-ups covering different types of rating scales.

Recent Research Findings

1. Behavioral Expectancy Scales were developed for nine dimensions of behavior for department managers of a retail chain. When compared with Likert-type scales covering the same dimensions it was found that the Expectancy Scale ratings showed less leniency error, less halo error and less method variance. Seven of nine scales were satisfactory in terms of method variance and discriminate validity (Campbell, et al, 1973).

2. The USPHS has developed an Efficiency Report for evaluating the performance of physicians. The Report contains four kinds of performance evaluation methods. Findings based on the use of this report suggested that a scale containing 50 Forced-Choice tetrads and a 22-item checklist were more valid measures of performance than were rating scales for evaluating an intern's Job Proficiency and Personal Qualifications (Howell, et al, 1960).

3. Rating scales composed of Forced-Choice tetrads have been developed and successfully used to evaluate hospital attendants (Prien) and police officers (Smith, 1973).

4. In a student counseling program it was found that students (peers) could evaluate other students about as well as could the practicum supervisor assuming that the evaluating was done using a well-structured scale -- the Rating Scale of Counselor Effectiveness. However, student ratings were significantly more lenient than were counselor ratings and therefore had to be adjusted downward by a constant factor before they could be used to evaluate other students (Friesen and Dunning, 1973).

5. In a student counseling program it was demonstrated that ratings obtained from coached clients, trained in objective rating techniques, were more consistent with the ratings of counselor educators when evaluating the counseling effectiveness of counselor trainees, as compared to the same ratings obtained from noncoached clients using the same criterion (McIlvaine, 1972).

6. In a study concerned with rating the "goodness" of tennis strokes it was found that beginning tennis students were more accurate in identifying movements which were performed correctly than those which were performed incorrectly (Osborne, 1973).

7. A variety of research cited by Campbell et. al. indicates that raters are reluctant to make judgmental ratings, but are willing to accurately report whether or not they have observed a person performing a specified act on their job (Campbell, et al, 1970).

BEHAVIORAL (EXPECTANCY) SCALES (BES)

Description

Countless efforts have been made to develop rating scales which, when coupled with behavioral observation, can provide a device for measuring performance. "Unfortunately, there is a rather long and sorry history of gallant but essentially failing efforts to develop rating scales for observing and measuring job behavior (Campbell, 1970)." Recently Campbell and others have used a combination of methodologies to develop what they claim to be an improved type of job behavior rating scale. There is some evidence to suggest that they have indeed done so.

In a study reported in 1968 Dunnette, Campbell and Hellervik developed a set of job behavior observation scales for store managers of a large retailing chain. The development of these scales was based on the combination of two methods -- the Critical Incidents Method (Flanagan, 1954) and the Retranslation Method as reported by Smith and Kendall (1963).

A brief description of the procedures for developing a B.E.S. is as follows:

1. For a particular job position, knowledgeable content experts are asked to provide critical incident descriptions of unusally good or unusually poor behavior.

2. The descriptive statements are analyzed and sorted by measurement specialists according to the job dimensions they seem to represent.

Rough definitions are provided for each dimension.

3. The content experts then are asked to sort the statements into the job dimension categories developed for them. A discussion ensues regarding how the dimensions might better be defined, what if any new dimensions are required, and so on.
4. The content experts then sort the statements for a particular dimension on a nine-step "effectiveness" dimension ranging from grossly ineffective to highly effective. As required additional behavioral descriptions are prepared to define those scale points not defined adequately by the original set of critical incidence statements.
5. For certain incident statements the experts will not agree on where they should be located on the nine-point "effectiveness" dimension. These statements are discarded. What remains is a set of nine-point dimensions each point defined by one or more behavioral statements.
6. The statements for all dimensions then are intermixed and given to a second group of content experts. Their task is to sort the statements into piles each pile representing a dimension of behavior. Once this is done they define the dimension then sort the statements in each dimension on a nine-step "effectiveness" dimension. Hopefully, the second group of content experts will arrive at the same set of dimensions and scale-point definitions as did the group that originally developed the scales. This is the "retranslation" method as described by Smith & Kendall, 1963).

7. Those dimensions of behavior not identified by both groups are eliminated from the assessment battery. For the remaining scales a consensus obtained by combining the sorting results of both groups can be used to select behavioral definitions of the scale points.

Campbell (1970) cites many advantages to Behavioral Expectancy Scales to include: (a) they employ job-related language; (b) they are based on a detailed analysis of what is essential for acceptable performance; (c) they force the rater to respond in terms of what behavior actually was exhibited on the job.

Obviously such scales are very time consuming to develop. In addition they seem to be situation specific. That is, scales developed to rate physician behavior in a clinic probably would not be applicable to rating the behavior of a private practitioner. Also, the approach tends to subdivide job behavior into many parts each of which must be assessed by its own scale. This greatly increases scale development and administrative costs.

Judgments About Technique

1. Reliability. Not reported.
2. Validity. Based on multi-trait, multimethod analysis of the degree to which a trait has construct validity; .30s to .60s reported.
3. Current State of Refineness. Quite well refined. Techniques for developing scales are based on two established procedures -- the Critical Incidents Method and the Retranslation Method.
4. Feasibility of Use.
   a. Cost. No information reported. Would seem to depend on

scale development costs, which probably are quite high.

b. People Requirements. Two or more raters each of whom have had an opportunity to frequently and recently observe job-related performance of a ratee. Raters should receive a small amount of training.

c. Test Environment Requirements. None. Raters judge performance as they observed it in a natural work setting.

d. Suitability for Nation-Wide Testing Program. Technique requires a multi-person work setting. Would seem suitable for assessing interns and residents. On some dimensions (ethics, inter-personal communication, etc.) technique might be used to assess a physician who practices mostly in his own office. Use peer or subordinate ratings.

5. Appropriateness for NBME R&D Support

a. Refinement Potential. Good potential. Research now being done to improve technique.

b. Adaptability to MD Assessment. Yes. Adaptable to all aspects of behavior where critical incidents can be used to define a dimension of behavior, and, where the work setting is such that there exist other persons who could serve as raters.

c. Potential for Assessing Non-Cognitive Processes. Quite high. For example, can be used to assess ethical behavior to the extent that judges can agree on what are examples of good, bad and indifferent ethical behavior.

FORCED-CHOICE RATING SCALES

Description

Considerable dissatisfaction has been expressed with conventional rating scales. In particular, they are subject to the "halo effect," leniency effect, and can easily be rigged by the rater. Forced-choice scales try to eliminate these possibilities by disguising the direction of the desired or "socially acceptable" response.

Using a forced-choice rating scale, a rater is asked to select two statements which describe specific examples of behavior. Typically, the rater is presented with four statements and asked to identify those two statements which best describe the field performance of the ratee. By statistical design, two of the statements are descriptive of outstanding job performance. The other two descriptions appear to describe outstanding performance but in fact do not.

The statements are derived by having a group of job incumbents record written examples of what they believe to be outstanding performance. These statements are then rated by a group of supervisors in terms of the extent that they do describe behavior which distinguished between good and not-so-good performers. Those statements that are descriptive of outstanding performance are determined by statistical analysis of supervisor evaluations of each statement.

The rating scale is composed of tetrads of questions, all of which seem to be socially acceptable but only two of which have been judged to be descriptive of outstanding performance.

Some of the advantages of such scales include:

1. They are based on actual descriptions of job-relevant behavior.
2. They cannot be easily rigged.
3. They require the rater to rate on the basis of actually observed job behavior.
4. Can be used to obtain peer ratings.

Judgments About Technique

1. Reliability. Not reported
2. Validity. Concurrent validities of .25 to .60 reported by Smith & Schau (1973). Concurrent validities of .61 to .83 reported by Howell, Cliff and Newman (1960).
3. Current State of Refineness. Fairly well defined. Procedures for collecting descriptions of good and poor performance and of relating these descriptions to supervisor judgments probably need further refinement. Criterion development probably needs refinement.
4. Feasibility of Use.
   a. Cost. No specific information reported but time involved in development of a scale does not seem excessive. Should be less costly than the development of Behavioral Expectancy Scales.
   b. People Requirement. Need content experts to provide statements; supervisors or other content experts to judge relation of each statement to good and poor performance; two peers to provide ratings.

c. Test Environment Requirements. None. Ratings based on performance observed in natural work setting.

d. Suitability for Nationwide Testing Program. Use of technique requires a multi-person work setting. However, scales are more applicable than Behavioral Expectancy Scales for use by peers and subordinates.

5. Appropriateness for NBME R&D Support.

a. Refinement Potential. Good potential. Should check with researchers who had developed scales for use with medical personnel.

b. Adaptability to MD Assessment. Yes. USPHS has done this. See Howell, Cliff and Newman (1960) reference.

Description

In 1972 Blanz and Ghiselli reported on the use of a new merit rating method, the Mixed Standard Scale. The method was first developed in Finland and recently field-tested at the Kaiser Aluminum & Chemical Company. The scale attempts to "minimize the common errors of rating (halo and leniency effects) and to improve the reliability of ratings." The scale is a "hidden" scale in that it attempts to obscure from the rater the nature of the traits being rated. The scale "permits an evaluation of the reliability with which each individual is rated, each scale rates, and each rater rates". Finally, the scale allows one to rate a person on a number of traits, each presumably related to the criterion behavior for good job performance.

The scale reported on by Blanz and Ghiselli was developed to assess middle-level managers. Eighteen traits were selected which were judged to be related to good managerial performance -- general appraisal, efficiency, diligence, initiative, and so on. For each trait a three-point scale was visualized and a behavioral definition developed for each point. For example, the definitions for the "initiative" trait were:

"I He is a real self-starter. He always takes the initiative and his superior never has to stimulate him.
II While generally he shows initiative, occasionally his superior has to prod him to get his work done.
III He has a bit of a tendency to sit around and wait for directions."

The final scale consisted of 54 statements, three for each of the eighteen traits to be assessed. On the rating form the 54 statements

were listed in random order. The rater responded to each statement by recording:

1. If the statement "fits the ratee"
2. If "the ratee is poorer than the statement"
3. If "the ratee is better than the statement"

The report authors show how the three graded statements for each trait can be converted into and scored as a seven-point scale. Rater consistency is easily determined, and a "rating accuracy" score can be developed for each trait scale and for each rater.

Judgments About Technique

1. Reliability. Not reported. Study was concerned with rater consistency and accuracy.
2. Validity. The 18 scales were factor analyzed and four factors were obtained. The relation between the factor scores and a measure of managerial success (a simple adjective rating scale) yielded correlations ranging from .34 to .45.
3. Current State of Refineness. Technique is refined, but to apply it to medical practitioners would require the development of a completely new set of traits.
4. Feasibility of Use.
   a. Cost. No information reported. Cost would depend on amount of time devoted to the identifying of traits to be covered by the scale, and the development of definitional statements for each trait.
   b. People Requirements. Same as for Forced-Choice scales.

c. Test Environment Requirements. None

d. Suitability for Nationwide Testing Program. Same as for Forced-Choice scales.

5. Appropriateness for NBME R&D Support.

a. Refinement Potential. Unknown, because method reported on in only one study.

b. Adaptability to MD Assessment. Yes, to the extent that there would be persons available to serve as raters, persons who have an opportunity to observe physician performance. Has the apparent advantage of assessing a number of traits with a relatively easy and quick-to-administer assessment instrument.

REFERENCES: RATING SCALES

Behavioral Expectancy Scales

1. Campbell, J. P., et al. The Development and Evaluation of Behaviorally Based Rating Scales, Journal of Applied Psychology, 1973, Vol. 51(1), 15-22.

2. Fogli, L., et al. Development of First-Level Behavioral Job Criteria, Journal of Applied Psychology, 1971, Vol. 55(1), 3-8.

3. Gluskinos, R. & Brennan, T. F. Selection and Evaluation Procedures for Operating Room Personnel, Journal of Applied Psychology, 1971, Vol. 55(2), 165-169.

4. Lawler, E. E., III, The Multitrait-Multirater Approach to Measuring Managerial Job Performance, Journal of Applied Psychology,1967, Vol. 51(5), 369.381.

5. Smith, P. C., & Kendall, L. M. Retranslation of Expectations: An Approach to the Construction of Unambiguous Anchors for Rating Scales., Journal of Applied Psychology, 1963, Vol. 47 (2), 149-155.

6. Zedeck, S. & Baker, H. T. Nursing Performance as Measured by Behavioral Expectation Scales: A Multitrait-Multirater Analysis, Organizational Behavior & Human Performance, 1972, Vol. 7 (3), 457-466.

Forced-Choice Rating Scales

1. Howell, M. A., Cliff, N., & Newman, S. H. Further Validation of Methods for Evaluating the Performance of Physicians. Educational and Psychological Measurements, 1960, Vol. 20(1), 69-78.

2. Howell, M. A. & Newman, S.H. Prediction of Supervisors' Ratings of Medical Intern Performance. Psychological Reports, 1971, Vol. 29(3, Part 2) 1325-1326.

3. Prien, E. P. Measurement of Performance Effectiveness of Hospital Attendants, American Journal of Mental Deficiency, 1972, Vol. 77(1), 21-25.

4. Smith, David H. & Schau, E. Forced-Choice Performance Rating for Policemen. Proceedings of the 81st Annual Convention of the American Psychological Association, Montreal,Canada, 1973, Vol. 8(2), 987-988.

Mixed Standard Scales

1. Blanz, F. & Ghiselli, E. D. The Mixed Standard Scale: A New Rating System, Personnel Psychology, 1972, Vol. 25(2), 185-200.

2. Egan, D. M. Executive Performance Appraisal: An Application of Non-Metric Multidimensional Scaling. Proceedings of the Academy of Management, 1971, 75-80.

Textbooks

1. Campbell, J. P., Dunnette, M. D., Lawler, III, E. E. and Weick, Jr., K. E. Managerial Behavior, Performance and Effectiveness, New York: McGraw-Hill Book Co., 1970.

2. Thorndike, R. L. and Hagen, E. Measurement and Evaluation in Psychology and Education (3rd Ed.), New York: John Wiley & Sons, Inc., 1969.

Recent Representative Studies

1. Dunnette, M. D. Managerial Effectiveness: Its Definition and Measurement. Studies in Personnel Psychology, 1970, Vol. 2(2), 6-20.

2. Friesen D. D. and Dunning, G. B. Peer Evaluation and Practicum Supervision, Counselor Education and Supervision, 1973, Vol. 12(3), 229-235.

3. McIlvaine, J. F. Coached Clients as Raters of Counseling Effectiveness, Counselor Education & Supervision, 1972, Vol. 12(2), 123-129.

4. Osborne, M. M., and Gordon, M. E. An Investigation of the Accuracy of Ratings of Gross Motor Skill, Research Quarterly, 1973, Vol. 43(1), 55-61.

5. Taylor, J. B., Haefele, E., Thompson, P., and O'Donoghue, C. Rating Scales as Measures of Clinical Judgment: II. The Reliability of Example-Anchored Scales Under Conditions of Rater Heterogeneity and Divergent Behavior Sampling, Educational and Psychological Measurement, 1970, Vol. 30(2), 301-310.

PERSONNEL EVALUATION SYSTEMS AND APPROACHES

Overview

Most industrial concerns periodically evaluate the performance of their employees. Performance evaluation provides the data upon which to base decisions about salaries, promotions, further training and so on. A description and evaluation of many of these industrial evaluation procedures appears in an excellent review published in 1970 by Campbell, et al. This text reports on findings based on "a comprehensive search for the available knowledge concerning the identification and enhancement of managerial effectiveness." Anyone interested in evaluation procedures as practiced in industry should study this text.

Unfortunately, after reviewing industrial practices concerning the definition and measurement of managerial effectiveness, Campbell was forced to conclude that "it seems that we are confronted with an interesting and rather surprising paradox: in industry today, job behavior is not typically observed and recorded with any degree of care or precision. Instead, decisions about people are often made on the basis of global impressions formed from job-irrelevant factors reflecting society stereotypes of success rather than on the basis of factors judged to be directly relevant to realizing broad organizational goals and outcomes."

Rating scales form the basis of most job performance evaluation approaches used in industry. Not surprisingly therefore, recent attempts to improve industrial evaluation systems have concentrated on developing better rating scales or developing more sophisticated

procedures for analyzing the data obtained from rating scales. In particular, industrial psychologists have been concerned with "(a) how to better identify appropriate criterion behavior; (b) how to more accurately identify the dimensions of behavior which should and can be assessed by rating scales; (c) how to define more precisely the points on a rating scale; (4) how to assure that those who use rating scales are properly trained in their use; and (3) how to assure that raters have adequate opportunity to observe those whom they are asked to rate."

Various special characteristics of the Military Services make it possible for them to employ evaluation practices which are less feasible to use in industry, and probably cannot be used to evaluate private practitioners. These special Service characteristics include:

1. All service personnel have supervisors who must evaluate their subordinates on an annual basis.
2. Officers and senior enlisted personnel continually attend new training programs and undergo evaluation as part of that training.
3. Both enlisted and officer personnel periodically take part in field maneuvers and tests. This requires them to practice for these exercises and their competency as individuals is usually evaluated during the exercise.
4. Enlisted personnel take a written proficiency test each year and are required to study for this test.
5. Officers are rotated to a new job every two or three years and thus are constantly being evaluated by a new supervisor.

Officers and enlisted personnel are constantly performing in the presence of others and thus are subjected to pressure, not too subtle at times, from their peers and supervisors to improve upon poor performance. Moreover, being constantly subjected to training exercises, incompetent personnel eventually are identified.

There are official Service-wide systems for evaluating officers and enlisted personnel. For the most part, these systems are based on the use of rating scales and descriptive efficiency reports. Administratively, these systems are useful. It is doubtful whether they generate valid evaluations.

The Military Services are constantly evaluating their personnel, but do this in a training setting. The Service's system of continuing education assures that the performance of personnel will be assessed from time to time. In recent years, teachers have adopted a strategy which has some similarity to the military's continuing education approach. This will be discussed under the heading of Competency-Based Education and Certification Procedures.

This section will conclude with a description of the multitrait-multi (method) rater approach, a statistical analysis approach to the interpretation of the validity of rating scales and other assessment methods.

PERFORMANCE EVALUATION SYSTEMS

Description

Performance Evaluation is the label applied to a set of performance-evaluation systems used throughout business and industry. Such systems are used to regularly and formally evaluate each person in an organization. These evaluations may be used to select a person for another job within the organization, to decide on the size of the salary increase to award a person, and/or to provide a record of the past year's performance of a person.

Performance Evaluation systems usually include a face-to-face meeting of an employee with his supervisor. At this meeting they review the substance of the supervisor's evaluation report for the employee. This meeting can easily become a confrontation. For this and other reasons, many companies and business educators are becoming increasingly disenchanted with Performance Evaluation systems. In its stead, they are attempting to develop other evaluative procedures which are based more on self-appraisal and outcome evaluation.

Most organizations have a somewhat unique Performance Evaluation system. However, most systems employ two or more of the following assessment techniques.

1. Essay appraisal
2. Graphic rating scales
3. Field review
4. Forced-choice ratings
5. Critical incident appraisal
6. Management-by-objectives approach
7. Work standards approach
8. Ranking methods
9. Assessment centers

Each of the above techniques is rather widely used in industry, and will be reviewed individually later on. Here we wish to review Performance Evaluation as a system.

Critics of Performance Evaluation, supported by some evidence, claim that people who receive honest but negative feedback, often performed more poorly after a performance appraisal interview. Proponents of the method concede this but point out that the proper use of techniques 2 through 9 above can prevent this.

Critics contend that with most Performance Evaluation systems, when one member of an organization is given a high evaluation this must be offset by giving someone else a poor evaluation. Proponents concede that this often occurs, but contend that ways can be found around this application problem.

Both critics and proponents agree that an appraisal review should not concentrate on the faults of an employee. Rather, the preferred technique is to get the employees to set their own performance goals for the next six months or so. During the next months, the supervisor is free to help the employee gain these goals. The subsequent appraisal interview concentrates on why the goals were or were not attained, what should be done about it, and what should be the employee's goals for the next six months. It is felt that these and similar techniques allow the employee to set, within limits, his own performance improvement goals, and allows the supervisor to take on the role of a helper rather than an evaluator.

Judgments About Technique

1. Reliability. Unknown. Little reported research on the system.
2. Validity. Unknown. Critics cite anecdotal evidence that the method is not valid. Proponents claim method can be valid if properly implemented.
3. Current State of Refineness. As an evaluation system, it is rather unrefined. The administrative procedures may appear very precise and the data processing procedures may be quite exacting. Unfortunately, the data are often suspect -- GIGO.
4. Feasibility of Use.
   a. Cost. Depends on techniques employed, and evaluation setting. Assuming resident physicians evaluating interns in a training hospital setting, the cost might not be too high.
   b. People Requirements. Something like one day of supervisor time for each person evaluated.
   c. Test Environment Requirements. This evaluation system is based on both direct and indirect observation and measurement in a natural work setting. Therefore, need a hospital or a physician's office, patients, etc. to provide a natural medical delivery setting.
   d. Suitability for Nationwide Testing Program. Not suitable unless some means can be found to provide test settings, obtain cooperation of supervisory personnel; also, means would have to be developed for providing assessment-oriented training to supervisory personnel.

5. Type of Behavioral Process Typically Assessed by Method. Can be and is used to assess all types of behavioral processes. However, system is usually employed to assess mediational processes, especially problem solving and decision making. Also, system is commonly used to assess interpersonal behavior. Seldom use system to assess motor processes. System can be, but seldom is, employed to assess performance in a cross-cultural setting. Assessment usually pertains to a "same-culture" setting.

6. Appropriateness for NBME R&D Support

   a. Refinement Potential. Long-range potential, but doubtful short-range payoff. Depends also on potential for refining individual techniques employed in Performance Evaluation systems.

   b. Adaptability to MD Assessment. Yes. Probably do this now in some hospitals and clinics.

   c. Potential for Assessing Non-Cognitive Processes. Quite high.

References

1. Dayal, I., Some Issues in Performance Appraisal. Personnel Administration, 1969, Vol. 32(1), 27-30.

2. Meyer, Herbert E., The Science of Telling Executives How They're Doing, Fortune, January 1974, 102-112.

3. Oberg, Winston, Make Performance Appraisal Relevant, Harvard Business Review, 1972 (Jan.-Feb.), Vol. 50, No. 1, 61-67.

4. Scharinger, D. H., Performance Appraisal: A Means or an End, Training and Development Journal, 1970, Vol. 24(3), 52-53.

5. Thompson, Paul H. and Dalton, Gene W., Performance Appraisal: Managers Beware, Harvard Business Review, 1970 (Jan.-Feb.), Vol. 48, No. 1, 149-157.

MILITARY OFFICER EFFICIENCY REPORTS

Description

All of the military services have developed procedures for evaluating officer personnel. Rating scales are used to evaluate: personal qualities, performance of duty factors; potential for promotion and for schooling and assignment potential. The rater provides also an overall rating of the ratee's "overall value to the service." The rater, by way of justifying his ratings, provides brief narrative comments to amplify certain of his ratings.

An Officer Efficiency Report is completed at least on an annual basis by a person's immediate supervisor; the ratings are reviewed and endorsed by a higher-level supervisor.

As stated in Army Regulation 623-105, "Efficiency reports are of such far-reaching importance to the Department of the Army and to the individual officer that the greatest care must be exercised in their preparation." They are, in fact, so important that, unless an officer is a complete dolt, he receives at least "above average" ratings on all scale dimensions.

Officer Efficiency Reports are subject to severe distortion by both halo and leniency effects. They apparently have a low but significant validity as determined by their relation to peer ratings and supervisor ratings. Undoubtedly, they do help to identify the incompetent officer, but that person would be identified without the use of an Efficiency Report.

Analysis of Narrative Evaluation Reports (ER)

Performance Evaluation systems employed by the military services and by some industrial firms include one or more narrative sections which contain evaluative descriptions and judgments about job incumbents. Reading these narratives, one can obtain an impression as to whether or not the person under consideration is, or is not, highly competent. Unfortunately, most such narratives, like letters of recommendation, tend to be filled with superlative statements.

As part of the Navy's enlisted personnel evaluation system, attempts have been made to develop new sources of information. One attempt dealt with the "Content Analyses of the Narrative Sections of Navy Performance Evaluation for Senior Enlisted Personnel." This study, one of the better of its kind, developed "29 categories for indexing the substantive content of the narrative text" of evaluation reports prepared for 225 senior enlisted personnel. The indexing categories were based on the development of an indexing vocabulary consisting of the words and phrases commonly used to describe performance in an evaluative fashion. Examples of such words are: above reproach, exceptional, above average, better, adequate, capable, declining quality, in need of, bottom and least. A group of judges was used to develop a five-category weighting scale and each indexing term was assigned to one of the five categories. By these procedures, terms such as "highest" and "ideal" were assigned a weight of five. Terms such as "lowest" and "less able" were assigned a weight of one.

The narrative portion of an ER then was examined in terms of the various evaluative indexing terms contained in it; the weight for each

term could be summed to arrive at an overall evaluation score. These scores were compared with the performance ratings which had been assigned to the persons under evaluation using other data. A fair amount of agreement was obtained.

The report authors concluded that "the in-depth content analysis methodology used in the study allows one to use narrative reports to discriminate between persons of different levels of competency."

The above conclusion can be compared with those of an Air Force study which found that the "word picture" of the Officer's Evaluation Report used by the Air Force cannot by itself serve as an effective discriminator of relative performance.

References

1. Coyle, H. E. & Gorman, C. D., The Utility of OER Word Pictures as Discriminators, USAF AFHRL Technical Report, 1970, No. 70-23.

2. Ramsey-Klee, D. M., A Content Analysis of the Narrative Section of Navy Performance Evaluations for Senior Enlisted Personnel, U. S. Naval Personnel and Training Research Laboratory Technical Report, 1971, No. 2-71.

ENLISTED PERFORMANCE EVALUATION SYSTEMS

Each of the Military Services has a rather elaborate Enlisted Performance Evaluation System. The efficiency reports and test scores generated by these systems serve many purposes. They force each enlisted person to engage in some type of continuing training program; they provide a way of ranking individuals so that those with the higher ranks can, if funds are available, be awarded extra pay in the form of "proficiency pay"; they provide relative rankings which then can be used to determine eligibility for promotion. Enlisted Evaluation Systems assess performance from a normative rather than from a criterion-based standpoint. In recent years more and more persons have questioned the validity of these systems.

Enlisted Evaluation Systems employed a variety of assessment approaches -- annually prepared efficiency or performance reports; annually administered paper and pencil tests covering one's occupational specialty and/or general military skills and knowledges. The Efficiency Reports usually contain two parts -- a rating scale and a "word picture" (a verbal description to support the judgments recorded on the rating scale). This information is placed in one's personnel file along with a variety of other information -- information about special citations and awards, graduate ranking in attended training programs, and so on. All this information is considered at promotion time.

Proficiency Testing occurs annually and is based on the administration of paper and pencil tests. The correlation between Proficiency Test scores and supervisor evaluation of proficiency usually is reported to be

in the range of .30 to .40. Proficiency Tests are used to determine who will become eligible for additional pay.

Enlisted Personnel promotion practices are fairly similar for all the services. At one time specially convened "promotion boards" selected persons for promotion on the basis of a subjective review of the content of their personnel file. To remove some of the subjectivity of "selection for promotion" the Air Force recently developed a "weighted Airman Promotion System" for enlisted personnel. Instead of relying exclusively on the judgments of a promotion board, one's eligibility for promotion is based on six factors which have been assigned different weights. The factors relate to:

a. Scores on two different tests -- a specialty test covering one's military occupation, and a general test.

b. Two scores derived respectively from time-in-service and time-in-grade.

c. A decoration score -- how many decorations, letters of commendation, and other special awards the enlisted person has received.

d. A subjective "performance rating" score derived from reading the descriptive job performance-related information contained in annual efficiency reports.

The Air Force "validated" the weighted Airman Promotion System by demonstrating that it could predict reasonably well those enlisted persons who would be identified as eligible for promotion by a promotion board. The predictive validity ranged between .47 and .56 depending on the specific Air Force occupational specialty under analysis.

Recently the Navy has been experimenting with an assessment technique based on the collection of critical incidence data. The experimental version of this technique has been developed and evaluated using electronic technicians as subjects (Rafaz, Siegel).

The job of an electronic technician was analyzed and eight critical job activity factors identified -- how to read manuals, how to use test equipment, and so on. A Job Performance Questionnaire then was developed. Using that Questionnaire supervisors recorded, for a particular technician, incidences of uncommonly good or uncommonly poor performance as observed during a specified period of time. Statistical procedures were employed to develop a weighted score representing "performance reliability" -- the degree to which one can rely on a person to perform satisfactorily his assigned tasks.

Performance reliability scores were compared with scores developed from a Task Performance Check Form, a performance acceptability scaling device currently used to rate technician performance. Correlations between the two sets of scores for various types of technicians ranged from .43 to .49. Siegel concluded that the Job Performance Questionnaire technique is probably the Navy's most valid technician assessment procedure.

The Navy's Fleet Ballistic Missile Weapons System has a rather elaborate paper and pencil testing system for assessing the proficiency of certain operators and maintenancemen of a submarine crew. It is a unique evaluation system because the testing program is used to evaluate operator and maintenance training programs. If large numbers of fleet enlisted personnel score poorly on certain portions of a test then it is assumed that the training program(s) they attended is deficient. To our

knowledge this is the only instance in the military service where test scores are used to evaluate training programs.

While at sea the crew of a submarine participates in a number of drills. This provides the opportunity for practice, additional training, identification of proficiency weaknesses, and so on.

Each submarine has two complete crews; while one is at sea the other is in a training status. The crews rotate every three months. The crew in training status participates in a number of testing exercises. Following these the entire crew, complete crew teams, and/or individuals receive remedial training as required.

As appropriate, individuals, teams and the entire crew receive training on the operation and maintenance of new or modified equipment. Also, they learn to perform in accordance with new operating procedures.

From the foregoing discussion it is obvious that the problem of assessing the proficiency of service enlisted personnel is quite different from those associated with the assessment of physician proficiency. Military officer and enlisted personnel who hold jobs which are critical to the success of a military operation are constantly assessed on the job (drills and alerts) and continually return to a student status for further training and assessment.

Imagine if you will a well developed residency program for physicians wherein they are continually being objectively and subjectively evaluated. Imagine now a recertification program which required a physician to retake an approved residency every three to five years. That's what the military services have, at least in principle.

Judgments About Technique

1. Reliability of Method. Often not reported. When it is reported it may range as high as .90.
2. Validity of Method. Reported concurrent and/or predictive validity ranges between .20 and .50 depending on the particular occupational specialty and job position under investigation. However, criterion behavior scores or ratings often have a doubtful relation to what seems to be required for successful job performance.
3. Current State of Refinement. Administrative procedures have been refined, but not so for the rating scales and tests employed.
4. Feasibility of Using Method
   a. Cost. Very low comparatively speaking; can prorate costs over large numbers of assessed individuals.
   b. People Requirements. Need raters, test developers and administrators.
   c. Test Environment Requirements. Natural setting for performance ratings; facilities for administering paper and pencil tests.
   d. Suitability for Nationwide Testing Program. Highly suitable in terms of administrative feasibility.
5. Type of Behavioral Process Typically Assessed by the Method. Paper and pencil tests assess cognitive processes; rating scales and efficiency reports can assess all types of behaviors.

6. Appropriateness for NBME Support

a. Refinement Potential. Rather low. Depends on extent to which individual parts of the assessment system (paper and pencil tests, rating scales, efficiency reports) can be refined.

b. Adaptability to Assessment of Physicians. General approach could br adapted. Would need a completely different set of tests and rating scales. Possibility of adapting critical technique to physician assessment as reported by Rafacz, Royle and Siegel merits additional study. In theory the procedure is an attractive one.

c. Potential for Assessing Non-Cognitive Processes. High. Can use approach to assess any type of behavioral processes. Also, the critical incidence approach might be used to assess any type of behavioral process.

References:

1. Koplyay, J.B. Field Test of the Weighted Airman Promotion System: Phase I: Analysis of the Promotion Board Component in the Weighted Factors System, Air Force Human Resources Lab., Technical Report 69-101, April 1969.

2. Koplyay, J.B. Field Test of the Weighted Airman Promotion System: Phase II: Validation of the System for Grades E-4 through E-7. Air Force Human Resources Lab., Technical Report 69-102, May 1969.

3. Rafacz, B.A. & Foley, P.P. Preliminary Results on the Evaluation of a Fleet Post-Training Performance Evaluation Technique. U.S. Naval Personnel Research & Development Lab., No. WTR-73-10, January 1973.

4. Royle, M.H., James, J. and Robertson, D.W. Situational Factors in Navy Enlisted Performance Evaluation. U.S. Naval Personnel & Training Research Lab., Technical Bulletin STB 72-8, March 1972.

5. Seigel, A. and Federman, P.J. Development of Performance Evaluative Measures: Investigation into and Application of a Fleet Post-training Performance Evaluation System. Applied Psychological Services, Inc., Wayne, Pa. 1970.

COMPETENCY-BASED EDUCATION AND CERTIFICATION PROCEDURES

Description

Competency-based education and certification refers to a strategy, a set of methods for attempting to ensure that the members of a profession are competent. Teacher education programs are actively developing competency-based programs; State agencies are actively developing competency-based certification programs. Some 31 States had plans underway as of 1971.

The general assumptions upon which competency-based education and certification are based are simple to state and extremely difficult to implement. The more salient assumptions are as follows:

1. Professionals should be certified on the basis of proven capability to perform specific tasks in a natural work setting.
2. It is up to certification agencies, in cooperation with all interested parties, to establish criteria by which to judge competency.
3. Once certification agencies have established certification criteria training programs will naturally change their emphasis to fit the criteria.
4. Various levels of certification should be established -- after completing an approved course of instruction; after practicing under supervision for a period of time; after advanced formal and on-the-job training covering skills not covered in undergraduate training.
5. The most feasible way to improve certification practices in the near future is to establish improved accreditation standards

for undergraduate and graduate programs, to include internship programs.

The State of Washington probably may have the most completely developed competency-based certification system for teachers. It assumes that an initial teaching certificate will be awarded after a student has completed an approved undergraduate program. Eventually all such programs will have to be competency-based. A program is competency-based when:

1. Specific competencies (behaviorally described training objectives) have been described and it can be shown that these competencies are relevant to good job performance.
2. Criteria for assessing competencies during training have been spelled out in detail and are judged to be feasible.
3. Assessment of student competency is based on the assessment of their performance.
4. Rate of progress through a course is based on demonstrated competencies rather than number of academic credits obtained.
5. The instructional program is designed to facilitate the development and evaluation of specific competencies.

A competency-based program is similar to a manufacturing assembly line in that: (a) the specifications of the end-product have been well defined; (b) procedures for maintaining quality-control have been established; and (c) there is some group of persons who from time to time examine and then "certify" that quality-control procedures are being followed and that product outputs do meet specifications.

In a competency-based program it is not absolutely necessary that students be taught to perform a prescribed list of tasks agreed upon by

a professional group. Rather, a competency-based program must describe accurately what it is that its graduates can and cannot do at the time of graduation. This gives the program the freedom to teach whatever it wishes. However, because its training objectives have been described in detail, members of a profession can easily criticize the program if it does not conform to acceptable views of what students should be taught.

The State of Washington program prescribes something akin to an internship and residency period. The recent graduate of a teacher program would practice under supervision for one year. They then would be certified to teach without supervision. If they wished to be further trained, they could, under the supervision of a "coordinating teacher" learn how to function as a "coordinating teacher" (a teacher who is trained to teach other teachers during their internship period).

Some of the procedures being established in the field of education seem to be patterned after medical education practices; if so this has not been mentioned in the literature. It does appear to be a movement of particular interest to the NBME because:

1. It is wrestling with all the problems associated with deciding who should be involved in accrediting and certification.
2. The emphasis seems to be on restructuring training programs so that you can accredit these programs on the basis of how precisely they describe the competencies they intend to teach to students and how carefully they assess students.
3. The necessity of some type of supervised internship is recognized, and some states (Washington among them) apparently are planning to develop programs to train "coordinating teachers."

Probably this will be followed by the development of procedures to evaluate the extent to which individual schools have the capability to conduct an internship program.

References:

1. Daniel, K.F. and Crenshaw, J.W. What has been and should be the role of State Education Agencies in the development and implementation of Teacher Education Programs (Both pre-and inservice)?: A review and analysis of literature. Prepared for the Bureau of Educational Personnel Development, Division of Assessment and Coordination, U.S. Office of Education, Order No. OEC-0-71-3315, Sept. 1971

2. Elam, S. Performance-Based Teacher Education: What is the State of the Art? Bull. of the Amer. Assoc. of Colleges for Teacher Education, 1971, 24 (9)

3. Turner, R.L. Levels of Criteria. In B. Rosner (Ed.) The Power of Competency-Based Teacher Education: Final Report of the Committee on National Program Priorities in Teacher Education. Washington, D.C.: National Center for Educational Research & Development, Office of Education, U.S. Department of HEW, July, 1971.

4. A Special Issue on Competency/Performance-Based Teacher Education. Phi Delta Kappan, Vol. 55, No. 5, January 1974.

THE MULTITRAIT-MULTIMETHOD (RATER) APPROACH

Description

A measurement method is valid to the extent that it measures what it purports to measure and does not measure what it does not purport to measure. Many behavioral measurement methods yield scores which correlate quite highly with an independent measure of similar behavior. That's good. These same methods may yield scores which correlate rather highly with measures of behavior which supposedly are unrelated to that which the method was designed to measure. That's bad.

The multitrait-multimethod (rater) approach is a set of analysis procedures which allows one to assess the construct validity of scores obtained from various measurement procedures. The approach is based on the assumption that when conducting research on measurement devices and methods one should use more than one method to measure more than one trait or category of behavior. By examining the various intercorrelations obtained one can determine: (a) if a particular method measures what it claims to measure, and (b) if the same method does not measure what it claims not to measure.

To describe the main features of the multitrait-multimethod approach let us assume that two rating scales have been developed respectively to assess behavior related to two of the nine major areas of physician clinical performance -- "history taking" and "treatment." Assume further that two different categories of observers resident physicians and nurses, will use the scales to assess the performance of interns.

The process of establishing the nine major areas of clinical performance was one of categorizing examples of behavior until one arrived

at nine separate categories which logically appeared to be different from one another. Different behavioral categories were constructed. Now we wish to establish the construct validity of these categories. This is done in part by assessing the "convergent" and "discriminant" validity of the instruments developed to measure behavior related to each of the constructed behavioral categories or traits.

If one measures the same performance by two different procedures the obtained scores should correlate rather highly with one another. This is evidence for convergent validity -- "validation is typically convergent, a confirmation by independent measurement procedures (Campbell & Fiske)."

If a measurement technique designed to measure trait A seems to measure trait A quite well, and, seems to measure an unrelated trait B quite poorly, this is an example of "Discriminant Validity." Discriminant validity is much more difficult to demonstrate than is "convergent validity." Perhaps that's why discriminant validity seldom is reported in the literature. It is based on attempts to show that a trait as measured is differentiated from all other traits. Thus, discriminant validity occurs to the extent that:

1. A trait measure correlates more highly with another measure of that same trait than with any other measure of a different trait using a different measurement method.
2. A trait measure correlates more highly with an independent effort to measure the same trait than with measures designed to get at different traits which employ the same measuring method.

Applying the above desiderata to the assessment of medical "history taking" and "treatment" ability we would expect:

1. Scores related to history taking ability as obtained from resident physicians would correlate significantly with scores obtained from nurses. This assumes that each group of raters used the same rating scale and had suitable opportunities to observe the interns at their work. Assuming the results to be as expected, this would be an example of convergent validity.
2. Scores obtained when residents and nurses used the history-taking scale to assess history-taking should correlate more highly with one another than:
   (1) scores obtained when the same raters used two different rating scales to measure history-taking and treatment behavior, respectively
   (2) scores obtained when different raters used the two different scales to measure history-taking and treatment behavior.

The above findings, if they occurred, would provide evidence of discriminant validity. Correlations related to discriminant validity seldom are reported in the literature. When they are, the results suggest that the measuring methods do not discriminate well.

The multrait-multimethod approach has not been employed to a great extent. However, those who report it seem convinced that it has a prominent place in the measurement of complex human behavior. Those who use the method typically find that their assessment measures have

convergent, but not discriminant validity -- the measurement scores correlate too highly with measures of unrelated(?) behavior.

In discussing the use of the approach Lawler made the following comments; they seem representative of those made by other users of the approach:

1. Use of the approach allows the researcher to obtain a more sophisticated understanding of his criteria. It gives him insight as to what are some of the dimensions which best account for the total behavior of a job incumbent, and provide an insight into what dimensions of behavior should be and need not be assessed.
2. When either one rater is used or when only one trait is rated the validity of the criteria must be accepted on faith. If insignificant results are obtained it is impossible to know whether the problem lies in the criterion or in the predictor.

After reviewing some of the recent work on peer and supervisor ratings, Lawler concluded that: it would appear that ratings with a good degree of discriminant and convergent validity might be generally obtainable where well-defined traits are used, where the raters have good knowledge of managers' job performance, and where appropriate rating scales are used.

References:

1. Campbell, D.T. & Fiske, D.W. Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. Psychol. Bull. 1959, Vol 56 No. 2, 81-105.
2. Gunderson, E.K.E. & Ryman, D.H. Convergent and discriminant validities of performance evaluations in extremely isolated groups. Personnel Psychology, 1971, Vol 24, No. 4, 715-724.

3. Lawler, E.E. III. The Multitrait-Multirater approach to measuring managerial job performance. Journal of Applied Psychology, 1967, Vol. 51, No. 5, 369-381.

4. Kavanagh, M.J., MacKinney, A.C., & Wolins, L. Issues in managerial performance: Multitrait-multimethod analyses of ratings. Psychological Bulletin, 1971, Vol. 75, No. 1, 34-49,

5. Zedeck, S. & Baker, H.T. Nursing performance as measured by behavioral expectation scales: A multitrait-multirater analysis. Organizational Behavior & Human Performance, 1972, Vol 7, No. 3, 457-466.

Measurement Category VII: BEHAVIORAL MEASUREMENT

Contents

	Index No.	Page No.
BEHAVIORAL TESTS		78
Overview - Psychomotor/Perceptual Tests	11	78
Listening	12	82
SITUATIONAL TESTS AND ASSESSMENT PROGRAMS		85
Overview		85
The Assessment Center	13	87
Leadership Assessment Simulations	14	91
Military Human Performance Assessment Programs	15	94
SYSTEMATIC OBSERVATION SYSTEMS		99
Overview		99
Sequential Analysis of Verbal Interaction (SAVI)	16	103
Flanders System of Interaction Analysis (FSIA)	17	107
CERLI Verbal Classification System (CVC)	18	111
Medical Instruction Observation Record (MIOR)	19	114
Notation of Proxemic Behavior	20	117
An Audio/Video Recording System for Pilot Training	21	120
Interviewing	22	123
PERFORMANCE TESTING TECHNIQUES	23	126
Overview		126
Simulation Techniques	24	130
Gaming	25	135
Role-Playing	26	139
In-Basket Exercise	27	142
Natural Setting Techniques	28	146
Adaptive Performance Testing	29	152

Overview - Psychomotor/Perceptual Tests

By behavioral tests we are referring to those measurement techniques which assess basic motor or perceptual skills that are necessary for the performance of more complex criterion behaviors. For example, finger dexterity skill is important for the performance of certain complex behaviors of pianists or surgeons. Tests which assess such enabling skills should predict to some degree performance capabilities of physicians on more complex behaviors. Such a prediction assumes that a task analysis of physician performance occurred and the required underlying skills were identified. The authors are aware of only one such analysis (Olson, 1972). It is possible that those individuals who have succeeded through medical school and are practicing physicians have met at least the minimal psychomotor and perceptual skill requirements. In these areas at least further testing of physician's abilities may be unwarranted. However, for purposes of this report, psychomotor and perceptual skills assessment will be discussed in general, with one perceptual area (listening) described specifically.

Physicians obtain information in a variety of forms. These could be visual stimuli in the form of microscopic specimens obtained at autopsies, tracings on an oscilloscope, numbers on a laboratory report, films of x-rays, or the appearance or movements of a patient. Aural stimuli may be sounds heard through a stethoscope, responses by the patients to various inquiries, etc. Tactual stimuli could come from a

physical examination of the patient. In order to develop valid as well as objective and reliable behavioral skills tests, exercises in the various sensory modalities could be developed which would assess the ability of physicians to interpret data arriving in these various forms (e.g., high fidelity tapes of sounds, video tapes or movies of examinations, etc.).

Such skills tests could be developed using an applicable taxonomy of these skills which is based upon a thorough behavioral task analysis of physician performance. Such an analysis should reveal the entire spectrum of enabling skills from basic motor skills to communicative, to decision-making requirements of physicians' jobs. A taxonomy of psychomotor skills that could be adopted for this purpose is presented in a recent book by Harrow (1972). This is one of the first complete in-depth studies of a classification system of behaviors existing within the psychomotor domain. The purpose of the book is to provide a taxonomy for the classification of observable movement behaviors, but perceptual behaviors are also discussed and organized. Psychomotor skills assessment of physicians could be based upon a classification scheme similar to the one promulgated in this document. It could serve as the basis for conducting a functional task analysis of physicians' jobs to determine what, if any, psychomotor skills assessment should take place before, during and following formal medical education.

Once the physician's criterion behaviors have been identified and enabling skills determined, then the relationships of skills tests with these criteria can be established. Until that happens, a discussion of

specific psychomotor tests is not warranted. Such tests are continuously being developed. Recently, computer-based techniques (Feurzeig, 1971) have been effectively demonstrated which can monitor the performance of complex perceptual and motor skills. Following a thorough job analysis of physician behavior, those psychomotor and perceptual skills which are relevant might then be assessed or monitored using such techniques. One such perceptual area is "listening" which will be discussed in the next section of this report.

Judgments About Technique

1. Reliability. Depends upon specific tests - ranging from .3 to over .8 in some recent tests of manual dexterity.
2. Validity. Validation of these tests, as with other assessment devices depends upon an investigation of their relationship to criterion behaviors. Establishment of these criteria still remains to be done.
3. Current State of Refineness. Depends upon specific tests. Most psychomotor tests have been developed for use with children.
4. Feasibility of Use.
   a. Cost. Unknown. Depends upon tests.
   b. People Requirements. Usual test administration requirements.
   c. Test Environment Requirements. Special equipment or apparatus may be required for certain tests.

d. Suitability for Nationwide Testing Program. Depends upon specific tests.

e. Number of People Tested Each Year. Unknown.

5. Type of Behavioral Process Typically Assessed by Method. Sensory, motor, and/or perceptual processes.

6. Appropriateness for NBME R&D Support

a. Refinement Potential. Need a thorough behavioral job analysis more than specific test development and refinement.

b. Adaptability to MD Assessment. Depends upon range of psychomotor/perceptual skills determined to be assessable.

c. Potential for Assessing Non-Cognitive Processes. By definition, these are being assessed.

References

1. Harrow, Anita J., The Taxonomy of the Psychomotor Domain: A Guide for Developing Behavioral Objectives. New York: David McCabe Co., 1972.

2. Feurzeig, Wallace, Automated Instructional Monitors for Complex Operational Tasks, Final Report, Bolt, Beranek, and Newman, Inc., Report No. R-2272, Oct., 1971.

3. Olson, Carl J., et al. The Orthopaedic Training Study, Phase II 1968-1972. Final Report Supplement, Psychomotor Skills, Part B. Urbana, Illinois University Medical Center, 1972

Description

One perceptual area that deserves attention is that of listening skills assessment. Listening depends upon hearing, but the two terms are not synonymous. Listening is the process of interpretation. The term "auding" has been proposed for listening to linguistic material. Some studies in the literature have shown a generally strong and positive relationship between listening ability and scholastic achievement. There is also general agreement that teaching listening skills is important to effective communication and that these skills can be taught and measured.

In the industrial world, recent surveys have shown that a large proportion (50%) of the major corporations sampled have included the teaching of listening skills in their management training programs.

There are very few published tests of listening skills of which the one by Brown and Carlsen (1955) is the best known and most widely used. It is designed to be used at the college and post-college levels. The test is divided into five parts, each measuring an important listening skill.

In addition to this test, there are a number of unpublished tests of listening which have been constructed as parts of various studies. It may be possible to develop a listening test from various refined sub-tests now in the experimental literature. Such an effort appears warranted considering the lack of assessment devices in this area.

In recent years, HumRRO researchers have performed extensive research in the area of listening or "auding." As part of a project for the Army (named REALISTIC), HumRRO investigators conducted a series of studies of the "auding" requirements for learning various job positions. Research in the "auding" area is also being conducted at several universities.

Judgments About Technique

1. Reliability. Only reported data concerns both forms of the Brown-Carlsen test. Correlations of .78 - .86 have been reported, although these data are old.
2. Validity. The validity of the Brown-Carlsen test rests on the fact that it was written to test those skills which a jury of experts deemed crucial in listening. However, the validity of the test has been called into question. The problem of the validity of any test of a previously unmeasured skill is very difficult to solve. The fact that this test has received a wide and continued use over a long period of time is meager evidence of its "acceptance" by experts in the field.
3. Current State of Refineness. Relatively unrefined for purposes of physician assessment.
4. Feasibility of Use
   a. Cost. Published, printed tests are inexpensive. If presented in taped version, tapes would add some to the cost, but if multi-channel tape playback units were available - test costs, per se, would be low.

b. People Requirements. Without tapes, "speakers" to present test would be required. This approach is not recommended.

c. Test Environment Requirements. For standardization purposes, tape-recorded presentations could be prepared - requiring tapes, playback equipment, earphones, etc.

d. Suitability for Nationwide Testing Program. Suitable if proper tests are developed.

5. Type of Behavioral Process Typically Assessed by Method

Perceptual and communication processes.

6. Appropriateness for NAME R&D Support

a. Refinement Potential. Definite potential for developing specific relevant tests of listening and/or "auding" proficiency of physicians.

b. Adaptability to MD Assessment. Considerable listening and "auding" behaviors seem basic to physician communicative and diagnostic capabilities. If criteria are established, such instruments could then be validated.

c. Potential for Assessing Non-Cognitive Processes. By definition, this is what occurs.

References

1. Brown, J. I, and Carlsen, G. R., Brown-Carlsen Listening Comprehension Test. Harcourt, Brace and World, 1955.

People and Places to Visit

Dr. Thomas G. Sticht, HumRRO Western Division, Carmel, Calif.

SITUATIONAL TESTS AND ASSESSMENT PROGRAMS

Overview

An endless variety of situational tests have been developed to assess human performance. Many of these are commonplace. We obtain a drivers license after demonstrating our ability to drive. We hire a secretary after giving her a battery of job-related tests.

Many situational tests are used for selection and classification purposes. Industry and the Military Services employ many of these. Many other types of situational tests are used to assess performance during or at the end of training. The Military Services employ extensively this type of test.

For a variety of reasons neither industry nor the military makes much use of situational tests developed specifically to assess the performance of experienced practitioners. Certain military tests and exercises come close to doing this. However, these can be considered as part of the military's "continuing training" program -- their primary purpose is to provide a vehicle for the identification of performance deficiencies which then can be corrected via subsequent training.

In recent years, considerable interest has been shown in the development of personnel assessment centers. These centers are used to evaluate the potential of senior technical personnel and first-level supervisors for assuming middle-level managerial positions. While these centers currently are used for selection purposes, it would appear that they might be usefully employed to assess the capability of experienced practitioners.

AT&T and other organizations have developed sophisticated assessment centers. However, these centers seldom employ sophisticated simulations of job related situations. Recently the U.S. Army developed a center for assessing the leadership potential of senior non-commissioned officers and junior officers. As part of this effort, HumRRO developed three rather extensive simulation exercises. A description of how such exercises are deve. ped may be of interest to the NBME.

What follows now are two descriptions re ted to the assessment center approach. Following this is a general review of situational tests and assessment programs as utilized by the Military Services. Finally, the assessment philosophy of the Military Services is discussed. It is a philosophy which promotes evaluation as a tool for improving performance. It has positive features which seem applicable to the assessment of any group of practitioners.

Description

Assessment Centers are used by many industrial firms to identify persons who have managerial potential. Such centers have the potential for providing the environment within which experienced managers could be assessed. Currently, they are not used in this fashion.

The concept of an Assessment Center is patterned after the centers developed during World War II to assess one's potential to become an intelligence agent (OSS assessment program). During a three or four day period, a group of persons are observed and evaluated under controlled conditions. Each person solves problems individually, participates in various team efforts, may be called upon to prepare and present a lecture, may be asked to prepare a report, and so on. In addition, each assessee usually is interviewed in-depth regarding his present and future work interests, present work deficiencies and plans for correcting them, and so on. A typical center program for assessing potential managers might be as follows:

1. Monday Morning. As a member of a four-person team, participates in a stock-trading game.
2. Monday Afternoon. Each assessee takes an in-basket exercise.
3. Monday Evening. Personal interviews are held early in the evening. Later in the evening, each person reads specially prepared financial material, then prepares a report discussing the pros and cons of acquiring another company.

4. Tuesday Morning. Assessee participates in a six person leaderless group discussion during which he supports a recommendation for a salary increase for a hypothetical employee.

5. ..... and so on ..... for three days.

The Assessment Center provides a "broad band" approach to the evaluation of executive potential. The assessee is observed in a variety of settings: management games, leaderless group discussions, case analyses, in-basket exercises, and interviews. In addition, he is observed over a considerable period of time.

The key ingredient of a successful Center operation is the trained assessor. This person is a content expert familiar with the exercises conducted at the Center. Assessor training is somewhat lengthy (3-5 days) and stresses: (a) an understanding of the exercises used in the Center and the types of behavior that they evoke, and (b) how to become a good observer, use whatever behavior recording devices are available, and prepare a final report of what was observed.

Prior to conducting a Center exercise one must identify the dimensions of behavior to be observed. These dimensions may be selected by soliciting suggestions from top management personnel, from consultants, or from both. The assessors must "discuss the specific meaning of each of these dimensions and are provided with rating forms that contain definitions of the meaning of each dimension."

Judgments About Technique

1. Reliability. Unknown. Not reported in research reports.

2. Validity. According to AT&T research, the Center assessment approach has reasonable validity. Six AT&T studies report correlation between assessment Center findings and subsequent ratings of field performance to range from .27 to .64

3. Current State of Refineness. Technique is fairly well refined. Increased use by the military and by industry should lead to further refinements.

4. Feasibility of Use.

   a. Cost. About $500 per person assessed, according to AT&T.

   b. People Requirements. One trained assessor per two assessees.

   c. Test Environment Requirements. Depends on exercises conducted at Center. Many management exercises are fairly easy and inexpensive to develop. Medically-related exercises could require an elaborate testing environment.

   d. Suitability for Nationwide Testing Program. Suitable if procedures and funds could be found to develop state or regional assessment Centers. Each industrial concern has developed its own Center.

5. Types of Behavioral Process Typically Assessed by Method. Commonly used to assess cognitive and inter-personnel communication processes. Seldom used to assess motor process. Assessment usually performed in a "same-culture" setting.

6. Appropriateness for NBME R&D Support

   a. Refinement Potential. Considerable potential depending on the degree to which one is willing to develop simulations of work environment.

b. Adaptability to MD Assessment. At present, can be readily adapted to the assessment of cognitive processes. Could be adapted (using actors, etc.) to assessment of interpersonal communication processes. Psychomotor skills also could be assessed.

c. Potential for Assessing Non-Cognitive Processes. Quite high.

References: Assessment Centers

1. Bray, D. W. and Grant, D. L. The Assessment Center in the Management of Potential for Business Management. Psychological Monographs, 1966, Vol. 80, No. 17.

2. Byham, William C. Assessment Centers for Spotting Future Managers. Harvard Business Review, 1970, Vol. 48, No. 4, 150-167.

3. Byham, William C. The Assessment Center as an Aid in Management Development. Training and Development Journal, 1971, Vol. 25, No. 12, 10-22.

4. Di Costanzo, Frank and Andretta, Thomas. The Supervisory Assessment Center in the Internal Revenue Service. Training and Development Journal, 1970, Vol. 24, No. 9, 12-15.

5. Huck, J. R. Assessment Centers: A Review of the External and Internal Validities, Personnel Psychology, 1973, Vol. 26, No. 2, 191-192.

6. Jaffee, C. L. Effective Management Selection: The Analysis of Behavior by Simulation. Addison-Wesley, Reading, Mass. 1971.

7. McConnell, J. H. and Parker, T. C. An Assessment Center Program for Multi-Organizational Use, Training and Development Journal, 1972, Vol. 26, No. 3, 6-14.

8. Slevin, Dennis P. The Assessment Center: Breakthrough in Management Appraisal and Development, Personnel Journal, 1972, Vol. 51, No. 4, 255-261.

LEADERSHIP ASSESSMENT SIMULATIONS

Description

Recently, the U.S. Army established an Assessment Center pilot program at Ft. Benning, Georgia. In many respects, this Center is similar to those used in industry. It does have one very unique feature, however. As part of the Center's assessment procedures three rather elaborate simulations are used to assess leadership potential. Simulations seldom are used in industrial assessment centers, probably because of the time and costs involved in their development and administration.

The three simulation exercises in use at Ft. Benning were developed by HumRRO Division #4. Their development has just been completed, thus, "the predictive validity of the simulations must await the accumulation of normative and criterion data." The simulations do, however, have considerable face validity, a necessity if one wishes to obtain the whole-hearted cooperation of those who must use the simulation. What follows now is the Summary and Conclusions of HumRRO Technical Report 73-21, Development of Leadership Assessment Simulations. Those interested in how such simulation exercises are developed and field-tested should find this report very informative.

PROBLEM

The U.S. Army has established an Assessment Center pilot program at Fort Benning, Georgia. The purpose of the program is to evaluate the feasibility of assessment centers for the Army. To accomplish this purpose, it was decided to develop and evaluate full procedures for evaluating the following levels of personnel: (a) student noncommissioned officers entering the Noncommissioned Officer Educational System (NCOES) senior program, (b) students entering the Infantry Officer Basic Course (IOBC) and the Branch Immaterial Officer Candidate Course (OCS), and (c) student officers entering the Infantry Officer Advanced Course (IOAC). The pilot plan encompasses development and evaluation of three complete assessment programs, one for each of the above levels.

HumRRO was requested to design and develop three leadership assessment simulation modules to be used in the assessment programs for the respective levels of personnel. Products of the research were to be (a) the "scenarios" of simulations appropriate for the three levels of personnel, to include the procedures and materials necessary to effectively conduct the simulations; (b) assessment procedures and instruments to be used with the simulations; and (c) procedures and materials needed to train assessment center personnel to conduct the simulations and use the associated assessment procedures.

APPROACH

The approach used was to (a) specify a set of leadership dimensions upon which to evaluate the personnel to be assessed; (b) develop simulations designed specifically to elicit behaviors relevant to the identified dimensions; (c) develop procedures and instruments for evaluating the relevant behaviors; and (d) develop materials suitable for training assessment center personnel to conduct the simulations and to perform the associated assessments.

RESULTS

Fourteen leadership dimensions were identified as appropriate for assessment and 11 were determined to be suitable for assessment by simulation. Three organizational simulations were developed to assess military personnel on the 11 dimensions.

Two of these simulations were designed within the context of combat operations. In the Senior Noncommissioned Officer Simulation (NCOES), the assessee performs the role of platoon sergeant and acting leader of a detached platoon engaged in stability operations. In the Senior Company-Grade Officer Simulation (IOAC), the assessee performs the role of the commander of a company that is part of an Infantry battalion engaged in stability operations.

Because of the lack of technical knowledge anticipated for entering OCS and IOBC students, the Junior Company-Grade Officer Simulation (OCS-IOBC) was designed within a non-combat context. In this simulation, the assessee performs the role of the leader of a platoon engaged in civil-disaster assistance in a city in the southeastern United States.

Each simulation lasts four or five hours, during which assessees experience a variety of problems under continually increasing environmental pressure. Their actions in

response to the problems are evaluated by assessors on the bases of preestablished criteria of effectiveness. During the course of the simulations, 111, 117, and 124 scores are obtained for NCOES, OCS-IOBC, and IOAC assessees respectively. Thus, a large number of scores with repeated measurements of most dimensions are obtained.

The final products were three separate "simulation modules," each of which contains;

(1) Full instructions and materials for conducting the simulation.
(2) Full procedures and materials for conducting the assessments associated with the simulations.
(3) Full instructions and materials for training center personnel, for conducting the simulations and using the assessment procedures and instruments.

CONCLUSIONS

(1) Organizational simulations possess the capability of effectively creating environments characteristic of complex hierarchical organizations and, through their emphasis upon performance in such environments, provide a contribution to the assessment process that cannot be obtained with any other technique.
(2) Like all assessment techniques, simulations are dependent for valid and reliable data upon observance of standard assessment practices, such as uniform conditions for all assessees and strict adherence to protocols concerned with administration of stimulus materials and scoring of responses.
(3) The effective use of simulations as assessment devices depends upon the competence of the staff responsible for conducting the simulations and making the assessments. This competence is achieved mainly through intensive training in use of the procedures and materials.

Description

Each Military Service employs several types of programs to assess personnel performance. One program is concerned with the formal evaluation of personnel by the use of efficiency reports. Features of this program were discussed under Measurement Category II: Appraisal by Others. In our opinion the efficiency and evaluation reports and attendant procedures used to implement this program are extremely deficient.

A second assessment program used by the military is not a formal program as such. Rather, it is a conglomerate of programs which collectively require each individual to participate continually in training programs, training tests, and alerts.

Each member of the Military Service must acquire and maintain two sets of skills. One relates to his ordinary job duties as performed during peacetime. These are acquired during training and are practiced and refined on the job. The second set relates to his wartime duties. These skills also are acquired during training but, unless in wartime, they are not practiced on the job. Rather, they are periodically assessed, practiced, and refined during special exercises -- field maneuvers, unit training exercises, and unit alerts.

Each military unit has a peacetime and wartime function. It is the wartime functions which are assessed during situational tests. These exercises serve a fourfold purpose -- they are a type of training exercise during which new members of a unit can learn how to function as a member of the unit; they allow experienced members of the unit to

practice and refine certain skills; they allow external observers to evaluate the proficiency of the unit; and they provide an opportunity for members of the unit to assess their own proficiency. The following descriptions illustrate the range of exercises conducted by the Military Services.

a. A missile firing team is transported to a firing range. They are asked to put the equipment for a surface-to-air missile unit into operation and to successfully fire two missiles. Twelve hours are given to complete the operation.

b. An Air Force base which is part of the Strategic Air Command is visited unannounced by an inspection team. The base commander is given 30 minutes to place the base on battle alert and to scramble 75 percent of his assigned aircraft.

c. The Captain of a naval vessel announces, without warning, that all personnel are to report to their battle alert duty stations within ten minutes. Various units of the ship then proceed to execute portions of an alert routine -- make guns ready for firing, get any ship-based aircraft airborne, secure the more vulnerable sections of the ship's hole, and so on.

d. An observer-evaluation team suddenly shows up and announces to the Captain of a B-52 that the observer team will go along on the about-to-commence training flight. The performance of all crew members will be evaluated during the flight.

e. An armored unit is scheduled to report to a maneuver range three months from now. Before that time they have to prepare their equipment for the maneuver, brief all their members

regarding the maneuver exercises they probably will perform, train as best they can new members to the unit, and practice whatever they can in whatever limited maneuver space is available to them.

From the foregoing descriptions it should be obvious that military testing programs make use of both planned and unannounced testing. In some instances (use of an observer-evaluation team to evaluate the crew of an aircraft) well developed rating scales and/or checklists may be used during the test. In many instances, however, "professional judgment" is used to decide whether or not unit performance was acceptable.

The most important feature of the military testing program is the amount of self-assessment that occurs during and after the test. Senior members of a unit often identify performance deficiencies in new unit members. Following the test they make very sure that these deficiencies are corrected.

After each test the senior members of the unit hold a debriefing session, sometimes with members of an evaluation team, for the purpose of identifying deficiencies in unit performance. A plan then is devised to assure that these deficiencies are corrected.

Most officers and senior non-commissioned officers realize that it is in their own best interests to identify and correct their individual deficiencies and those of the unit to which they belong. They are especially interested in doing this when they are responsible for all or portions of that unit's proficiency.

Evaluation as a Tool for Improving Performance. The Military Services more than any other organization can if they wish discharge a person for unacceptable performance of duty. Interestingly enough such an action rarely occurs on the basis of the formal assessment of performance. Rather, it occurs because the supervisors, peers, or subordinates to an individual become convinced that the person cannot properly discharge his duties and responsibilities. A series of counseling sessions, investigations and board actions then ensues often culminating in the discharge of the person under investigation.

What then is the purpose of the many test and evaluation exercises conducted by the Service? Their purpose is to provide an opportunity to learn, to practice, and to identify deficiencies which subsequently will be corrected. It is assumed that a military unit and its individual members can become acceptably proficient. If at any given time the unit and/or certain of its members cannot perform to minimum standards then it is assumed that a corrective plan of action will be developed and successfully executed. Given this as the prime purpose of performance evaluation, performance evaluation then becomes acceptable to the individuals involved. At some later time an inspection team may reassess the unit or at least examine the records which describe its recent remedial training activities. If it is judged that a member of the unit cannot or will not master his assigned duties then usually he is reassigned to another unit. If the individual's deficiencies are judged to be severe he may not be allowed to re-enlist.

An important parallel to that just described exists in industry. Job Performance evalaluation approaches have not worked well in industry

because they require a supervisor to critically evaluate his employees. As traditionally implemented the approach was negative. The supervisor was required to report on each employee's deficiencies and was supposed to discuss these deficiencies with each respective employee. This approach did not work. Supervisors were reluctant to criticize and to review these criticisms with employees; employees complained when they were never told about the contents of their efficiency report, and complained even more so when they were told.

In recent years industry has found that a performance evaluation system can be made to work when it is used as a vehicle for identifying employee deficiencies and developing plans to correct them. Sometimes this is called the "supervision by objectives" approach. It is discussed in some detail in Campbell's textbook (1970). (See references for Rating Scales.)

It would seem that the preceding approach to the use of assessment is the one most apt to be accepted by the medical community. From a statistical standpoint fifty percent of practicing physicians are below average in proficiency. However, the goal of any approach to physician assessment should not be to identify and weed out those physicians who have a sub-average proficiency. Rather, the goal should be to assist the large majority of acceptably proficient practitioners to become even more proficient.

Overview

Observation Systems are actually special instruments which have been developed to systematically codify behavior and generally require the use of trained observers. Analyses may or may not be sequential and observation may occur under real or simulated conditions. In some cases, the events have been video and/or audio taped. Taken together, systematic observation systems have been developed to categorize a wide variety of behavior in many different behavioral settings.

Some of the first systems developed were used to study the interaction process of small groups (Bales). Later on educators turned their attention to studying the nature of classroom interaction between teacher and pupils and applied similar techniques (Flanders). For the most part, interest has been confined to using the results for feedback and self-appraisal on the part of both teacher and students as well as for curriculum evaluation. No attempts have been made to relate interaction analysis to good or poor performance criteria as yet.

Any situation in which people interact in a face-to-face fashion is amenable to behavioral analysis. But, the system used for this purpose must include specifically defined behavioral categories appropriate to the situation under analysis. Although many observational systems exist, modification and/or creation of analysis categories is usually necessary when the system is to be used for a purpose other than originally intended. Associated with modifying an existing system or creating a new one are the goals of measuring the interaction in question.

That is, deducing the kinds of information needed to better understand the interaction process.

Classroom observational systems have recently been extended to industrial and counseling settings. The subject of observation may be one or more people and the number of observers required could be one or several.

One source (Mirrors for Behavior) describes in detail 93 Observational Systems, but there are many others in existence. The point is that many systems were developed for very specific purposes; other systems are merely expansions or attempted improvements upon those already developed. Most systems concentrate on categorizing verbal behavior, but some include non-verbal as well (gestures, eye movements,) etc.). Others categorize both.

Some attempts have already been made to assess the quality of communication of physicians using Bales Interaction Process Analysis. Several other techniques which might be considered for appraising physician-patient interpersonal processes, e.g., patient compliance, are reviewed separately following this discussion.

Judgment About Technique

1. Reliability - inter-rater reliability as high as the .90's has been reported.
2. Validity - unknown; no real validity data available.
3. Current State of Refineness - This is highly dependent upon the system used. There is a lot of information on how to refine observational systems.

4. Feasibility of Use

   a. Cost. Rather costly since trained observers are required and few people can be appraised at one time.

   b. People Requirements. Depending on the instrument, one or two teams of two persons each.

   c. Test Environment Requirements. Direct assessment would require a natural or possibly simulated physician-patient setting, most likely interview and/or examination.

   d. Suitability for Nationwide Testing Program. Not suitable unless ways can be developed to standardize the interpersonal situation and obtain high reliability among judges.

5. Type of Behavioral Process Usually Assessed by Method. Interpersonal processes such as clarity of communication, understanding and patient compliance.

6. Appropriateness for NBME R&D Support.

   a. Refinement-Potential. Problem is one of specifying criteria which are considered related to good and poor performance of physicians. Additionally, special categories of behavior would probably have to be added to existing systems.

   b. Adaptability to MD Assessment. Depends on validity of criteria.

   c. Potential for Assessing Non-Cognitive Processes. This can be done reliably if criteria are agreed upon.

References:

1. Mirrors for Behavior, An Anthology of Observation Instruments, Research for Better Schools, Inc. Philadelphia, Pa. 1967, Vols 1-14, Supplements A & B 1970.

2. Bales, Robert F. Interaction Process Analysis, Addison-Wesley Press, Inc. Cambridge, Mass. 1951.

3. Barro, Arlene R., Survey and Evaluation Approaches in Physician Performance Measurement. (Preliminary Version), Division of Educational Measurement and Research Association of American Medical Colleges, May 1972.

4. Handbook of Social Psychology, 2nd Ed., Vol. 2, Edited by Gardner Lindzey and Elliot Oronson, Addison-Wesley Publishing Company, 1968. (Chapter 13)

Sequential Analysis of Verbal Interaction (SAVI)

Anita Simon & Yvonne Agazarian

Description

SAVI is an observational technique designed to chart various stages of development of any group during its phases of problem solving. It provides both a theoretical and operational link between the affective and cognitive dimensions of communication. It is used to describe interaction processes and diagnose communication conditions necessary for any group to work more effectively. It can be used with any group size and allows for data collection about leaders and members other than school classes.

There are nine theoretical classes of verbal behavior (3 x 3 matrix); within the matrix are distributed 28 sub-categories of behavior. However, SAVI can be used with as few as four and as many as five dozen categories!

SAVI is based on the theory that every statement contains personal (affective) and topic (cognitive) components and the affective message must be clear and supported if the cognitive component is to be used. Therefore, cognitive statements with an affective message, that is implicitly negative, will less likely be used by the receiver. From this it follows that a leader who takes time to accept and work with the feelings of group members, accepting their cognitive messages in a climate of understanding, increases the probability that his own cognitive resources will be used by the group.

This system tries to analyze verbal communication in a way that reflects major patterns of both verbal and non-verbal interaction. SAVI classifies verbal behavior as either approaching or avoiding personal and topic information in the verbal message. Studying verbal interaction in this manner results in a simplified pattern of approach or avoidance of the communication of interpersonal information. All verbal acts are categorized as either approaching or avoiding the problem of what and how to communicate.

Judgments About Technique

1. Reliability. Intra-observer reliability was estimated when the same observer coded the same taped communication after a six-week interval; agreement with earlier coding was 86%. Inter-observer reliability was estimated when two observers coded ten 10-minute samples independently. Agreement ranged between 85% and 95% for the ten samples. Agreement for 2,807 codings was 71.4%.
2. Validity. Unknown. Although the theory seems to have some face validity, there is a need for it to be validated empirically. The SAVI system needs construct and predictive validation, category by category.
3. Current State of Refineness. For evaluative purposes, fairly refined if criteria specified.
4. Feasibility of Use for Physicians.
   a. Cost. Fairly costly. $100 - $300 a day per observer. Observers need to be trained. Also, if actors are used and video tapes made, the cost would be increased.

b. People Requirements. One observer for about each four people evaluated (estimated).

c. Test Environment Requirements. Testing can occur in natural or simulated work setting. Observation may be direct or indirect (taped).

d. Suitability for Nationwide Testing Program. If instrument is validated and criteria established empirically there appears to be no major problems.

5. Types of Behavior Process Being Assessed by Method. Interpersonal processes such as clarity of communication, understanding, and patient compliance.

6. Appropriateness for NBME R&D Support

a. Refinement Potential. Pretty good if criteria are established. May require some modification or change in categories.

b. Adaptability to MD Assessment. Yes, if criteria can be agreed upon.

c. Potential for Assessing Non-Cognitive Processes. Actually assesses both cognitive and affective components together. May be somewhat of a problem to separate out these components.

References:

1. Flanders, Ned A., and Simon, Anita, "Teacher Effectiveness," Mr. Robert L. Ebel (Ed), Encyclopedia of Educational Research 1970.

2. Mirrors for Behavior, An Anthology of Observation Instruments, Research for Better Schools, Inc. Philadelphia, Pa. 1967, Vol. III.

3. Simon, Anita, Use of Verbal Interaction Application I. Part III of a series of papers presented at the Annual Convention of American Association of Humanistics (undated).

Flanders System of Interaction Analysis (FSIA)
Ned A. Flanders

Description

FSIA is the most widely used classroom observation system. It can be coded directly by an observer in the classroom to obtain information about the social-emotional climate. Although the focus is on the teachers' behaviors which encourage or discourage student freedom, also included in the ten-category system are categories which classified student responses.

Usually, this instrument is used only in the classroom setting with both teacher and pupil as subjects of observation. However, FSIA has not been limited to studying a simple dyadic relationship of teacher and pupil since several students have been observed at one time. The Flanders system has primarily been used in research and training. No studies have been found in which evaluation or assessment was the purpose of the study.

Data collection is accomplished in a natural setting, and no special equipment is necessary. One trained observer can perform the required coding procedures.

The FSIA is concerned only with the socio-emotional climate of a teacher-student situation. Therefore, the system deals with the affective dimension only. Of the ten categories in the system, seven pertain to teacher verbal behavior, two to student verbal behavior and one to short periods of silence that are ambiguous or noisy. The seven teacher categories are divided into indirect and direct influence.

A breakdown of the ten category system is as follows:

1. Teacher Talk
    a. Indirect influence
        1) Accepts feeling
        2) Praises or encourages
        3) Accepts or uses ideas of student
        4) Asks questions
    b. Direct influence
        1) Lecturing
        2) Giving directions
        3) Criticizing or justifying authority
2. Student Talk
    a. Student talk-response
    b. Student talk-initiation
3. Silence or Confusion

Observers are usually trained on tapes of classroom interaction after they have memorized the categories. Then, they begin to use the system in the actual classroom. It is assumed that the system is designed to be totally inclusive of all possible events and a single event must be recorded in only one category.

Recent efforts have been made to describe and analyze non-verbal as well as verbal behavior using Flanders System as a base.

Judgments About Technique

1. Reliability of Method. Inter-rater reliability reported between .88 and .96 using Scott's reliability coefficient. A

Scott coefficient of 0.85 or higher is a reasonable level of performance.

2. Validity of Method. No validity data available.

3. Current State of Refinement. Flanders has recently expanded his system to include sub-classes for 8 of the 10 categories.

4. Feasibility of Using Method.

   a. Cost. About $100-$300 a day for each trained observer needed.

   b. People Requirements. One trained observer for each physician-patient interaction.

   c. Test Environment Requirements. Natural setting preferred but may be able to simulate or use audio-visual tape.

   d. Suitability for Nationwide Testing Program. This depends on the development of adequate data on validity as well as agreement on criteria for assessment.

5. Type of Behavioral Process Typically Assessed by the Method. Method might be adapted to assess physician-patient interaction, clarity of instruction and questions of physician, reinforcement by physician, compliance by patient.

6. Appropriateness for NBME Support

   a. Refinement Potential. Deletion and possible addition of categories probably required.

   b. Adaptability to Assessment of Physicians. Depends on whether suitable criteria of performance can be agreed upon.

c. Potential for Assessing Non-Cognitive Processes. Good. Existing categories reflect several facets of non-cognitive behavior of interpersonal relationships or social interaction.

References

1. Amodin, E. J. and Hough, J. L. (Eds.), Interaction Analysis: Theory, Research and Application, Addison-Wesley, Reading, Mass.

2. French, R. L. (U. of Tenn.) and Galloway, C. M. (Ohio State Univ.), A Description of Teacher Behavior: Verbal and Non-Verbal, Education Resources Information Center #028134.

3. Mirrors for Behavior, An Anthology of Observation Instruments, Research for Better Schools, Inc., Phila., Pa., 1967, Vol. II.

CERLI Verbal Classification System (CVC)
Cooperative Educational Research Laboratory, Inc.

Description

This instrument is the result of a sophisticated effort to combine the best from many existing category systems. A successful attempt has been made to simplify the coding procedures for easy interpretation of interaction. Essentially, it is a flexible two-dimensional system combining four process (how) categories: (seek, inform, accept and reject) with four substantive (what) categories: (memory, productive-critical thinking, expressed emotion and class management). This 4 x 4 matrix forms a sixteen category system.

Subjects of observation have been both teachers and students in the classroom setting. The results are usually fed back to the subjects so they may better understand their own interaction processes. So far, uses of the system have been confined to research and training; it has not been used for assessment purposes other than self-evaluation.

Only one observer is required for collecting and coding interactions which occur in a natural setting. No special equipment is needed. Category dimensions include affective, cognitive, social structure (role). Non-verbal behavior is not classified under this system.

As with all category systems, ground rules are provided when it is difficult to decide in which of two categories a communication belongs. There are "general," "process" and "substance" ground rules that cover most ambiguous interactions which might occur.

It is possible to record the sequence of communication as well as the identity of the speaker if this is desireable. The matrix appears adaptable to fulfilling a variety of needs which the user may have.

Judgments About Technique

1. Reliability. Unknown
2. Validity. Unknown.
3. Current State of Refineness. Fairly refined. There are specific rules to govern coding of ambiguous communications. Need empirical data on reliability and validity for assessment purposes.
4. Feasibility of Using the Method
   a. Cost. $100-200 per day for each observer used. Cost of training observers is minimal.
   b. People Requirements. One observer for each physician/patient interaction period.
   c. Test Environment Requirements. Natural or simulated setting, although audio-visual tape of interaction could be substituted.
   d. Suitability for a Nationwide Testing Program. Appears feasible. Observers should be easy to train.
5. Type of Behavioral Process Typically Assessed by the Method. Effectiveness of interpersonal behavior of physician in physician/patient interaction, particularly in affective domain.

6. Appropriateness for NBME R&D Support

a. Refinement Potential. Good. Modification of present system would be required to fit the needs of NBME.

b. Adaptability to Assessment of Physicians. Depends on establishment of valid criteria.

c. Potential for Assessing Non-Cognitive Processes. In its present form, non-cognitive processes are categorized as well as cognitive ones.

References:

1. Flanders, N.A., Interaction Analysis in the Classroom: A Manual for Observers, University of Michigan, Ann Arbor, Michigan, 1966.

2. Mirrors for Behavior, An Anthology of Observation Instruments, Research for Better Schools, Inc. Philadelphia, Pa. 1967, Vol. VIII.

3. The Role of the Teacher in the Classroom: A Manual for Understanding and Improving Teacher Classroom Behavior, Minneapolis, Minnesota, 1967.

Medical Instruction Observation Record (MIOR)

Description

MIOR was developed to assess medical instruction. This technique is actually a rating schedule and not a category system. It can be used to assess seven major dimensions of instruction; these include:

1. Attitudes toward individual differences.
2. Sensitivity to physical setting.
3. Attitudes towards students and patients.
4. Use of instructional materials.
5. Reaction to student needs.
6. Use of teaching methods.
7. Use of challenge.

A rating scale has been developed for each of the seven dimensions.

At the end of each scale, space is provided for comments and anecdotal evidence associated with the rating given.

Each of the scales has 22 points with detailed behavioral examples specified at five point intervals along the scale. This instrument is especially interesting because efforts have been made to translate evidence of attitude into behaviorally specified dimensions. Additionally, it focuses on the affective dimensions of medical education and away from cognitive and experiential variables.

A score is entered for each of the seven scales (profile) on the summary sheet of the observational record. Other types of information recorded on this sheet are: class level (years), class type (lab,

ward, classroom, etc.), teacher, school, department, topic, size of class, time observed, and name of observer(s).

Teachers at seven medical schools were studied using this instrument. Both medical and educational experts were trained as observers.

For scales 1, 3 and 5, it was generally found that an instructor's statements (verbal) as well as his actions and gestures (non-verbal) convey whether he favors student expression of individual opinion and communicates an atmosphere of friendliness or hostility which is felt by the student. The other scales pertain primarily to the instructional aspects of the situation.

Judgments About Method

1. Reliability. Inter-observer agreement; each observer with every other ranged from .81 to .96.
2. Validity. No validity data available at this time.
3. Current State of Refineness. Rating scales seem to be fairly well refined, although some are not applicable to physician assessment.
4. Feasibility of Use
   a. Cost. $300 - $500 for each observer per day.
   b. People Requirements. Several days for each observer used at each testing location.
   c. Testing Environment Requirements. Probably need live medical setting, although it may be feasible to videotape physician/patient interaction.

d. Suitability for Nationwide Testing Program. Depends on ability to recruit enough trained medical observers and on reaching agreement on valid performance criteria.

e. Types of Behavior Usually Assessed by Method. Could be adapted to assess attitudes of physician toward patient. Atmosphere created by physician in his interpersonal relationship to the patient, e.g., acceptance - rejection, friendliness - hostility, etc.

5. Appropriateness for NBME R&D Support

a. Refinement Potential. Fairly good, but will require some research to relate attitudes to effective performance.

b. Adaptability to MD Assessment. Might be able to use some of the existing scales such as 1, 3 and 5 to assess physician attitude toward the patient. Additional attitude scales would undoubtedly have to be constructed, however, to assess additional physician attitudes.

c. Potential for Assessing Non-Cognitive Processes. Appears to be high since several of the original scales were developed for this purpose.

References

1. Mirrors for Behavior, Vol. X.

2. Jason, H., A Study of Medical Teaching Practices, Journal of Medical Education, 1962, Vol. 37.

3. Buehler, J. A. and Trainer, J. B., Predictions of Medical School School Performance and its Relationship to Achievement, Journal of Medical Education, 1962, Vol. 37, 10-18.

Notation of Proxemic[1] Behavior - Edward T. Hall

Description

This is a system which has roots in anthropology. It is designed to describe the behaviors which constitute cultural patterns about which most people are not aware yet which profoundly affect their social interactions. For example, a person who gets "too close for comfort" triggers strong affective reactions in most people, even though they are not aware of having a comfort zone. Since the way people manage physical distances, posture, eye contact, gestures and so forth varies from culture to culture, this system provides a means of predicting which behaviors will more likely be welcome and those which will probably be distasteful to people of a given culture.

The categories which have been developed are strictly non-verbal and the dimensions of the system are psychomotor (body movement) and sociological structure (role) although an affective component (judgment of acceptability) is involved. Simple coding procedures permit data collection for the following categories:

a. Postural - sex identification
b. Sociofugal - physical orientation of subjects toward one another
c. Kinesthetic factor
d. Touch code
e. Retinal combinations
f. Thermal code
g. Olfaction code
h. Voice loudness scale

Up until now the system has been used primarily to study the interaction patterns of a variety of foreign cultures under a number of

[1] Study of microspace as a system of bio-communication.

different circumstances, including patient-therapist setting. However, there have been no attempts made to evaluate or assess social interaction using this instrument; all the work has focused on either research or training. Coding is performed on the basis of a time sample.

The Hall system has been described here because it represents an instrument of systematic observation which is based entirely on non-verbal behavior. It might be possible to adapt it to the assessment of physician-patient interactions on a cross-cultural basis, or even on a sub-cultural basis (socio-economic level).

Judgment About Technique

1. Reliability of the Method. Unknown
2. Validity of the Method. Unknown
3. Current State of Refinement. Existing categories well defined and symbols for coding are clear.
4. Feasibility of Using the Method
   a. Cost. $100 to $200 per day for each observer.
   b. People Requirements. 1 observer for each physician-patient interaction assessment.
   c. Test Environment Requirements. Natural setting or live simulation.
   d. Suitability for Nationwide Testing. Depends on availability of individuals who can be trained as observers, and on the validity and reliability of the instrument itself.
5. Type of Behavioral Processes Typically Assessed by the Method. Assessment of cross-cultural and sub-cultural physician-patient interactions.

6. Appropriateness of NBME R&D Support.

   a. Refinement Potential. Good. If proper criterion can be established for interpersonal behavior in physician-patient interaction.

   b. Adaptability to Assessment of Physician. Research required in this area.

   c. Potential for Assessing Non-Cognitive Processes. Very good if relationship between behavioral predictions and criterion can be established.

References:

1. Barker, R.G. & Barker, L.S., Behavioral Units for the Comparative Study of Cultures. In "Study of Personality Cross-Culturally," B. Kaplan (ED.), New York: Row Peterson, 1961.

2. Hall, E.T., A System for the Notation of Proxemix Behavior. American Anthropologist, 1963, Vol. 65, No. 5, pg 1003-1021.

3. Mirrors for Behavior, An Anthology of Observation Instruments, Research for Better Schools, Inc. Philadelphia, Pa. 1967, Vol. IX

AN AUDIO/VIDEO RECORDING SYSTEM FOR PILOT TRAINING

Description

The military services are constantly seeking improved methods for training and assessing pilots. During and prior to World War II student pilots were evaluated by pilot-observers who, after taking a check ride with the student, made a gross rating of student performance. In the absence of clear-cut standards and good recording procedures, each pilot-observer developed his own criteria for good pilot performance. The result was that the quality of new pilots was very uneven.

Since World War II, pilot training has become standardized, criterion behavior for standard flight maneuvers has been established, and an elaborate set of check lists has been developed to record pilot performance while in flight. These training and assessment method improvements have considerably increased the quality of graduate pilots.

Beginning in the early 1960's the Services, especially the Air Force, became interested in the use of audio-video recording systems as a means of obtaining a detailed record of student pilot performance. Such a system has not yet been fully developed although a fairly refined version was reported on in 1970. In its perfected state, the system would: (a) record pilot behavior during a practice or test flight; (b) record the readings on all significant control panel displays; and (c) continuously record the external environment as viewed by the student pilot. To accomplish all this, two or three TV cameras would be required.

As envisioned by the Air Force, the audio/visual recording system would be used as follows:

1. During pre-flight training sessions, a student pilot could view previous recordings of the specific mission he was about to fly. This would help prepare him for the mission.
2. During post-flight training sessions, the student could review his own performance both with and without the assistance of an instructor.
3. The recordings would allow an instructor to grade student performance after a flight instead of during a flight as is the current practice.

The preceeding recording procedures have obvious similarities to some of those employed with "Systematic Observation" methods. A number of studies have shown that ratings of performance or interpersonal interactions can be made either on the basis of live observations or the study of film or video tape. There are certain advantages to the use of video tape.

It is interesting to speculate on what might happen if one knew that his behavior was being video-taped for later evaluation. This technique is often employed in teacher education programs and during counselor training. We know of no instance where it has been systematically used to record and then to evaluate job incumbent performance. Under certain circumstances, it would be possible to do this. For example;

1. Surgical operations have been recorded in detail on film or video tape. Once recorded, the tapes can be used for either training or evaluative purposes.

2. Television cameras are used to provide surveillance in the public area of a bank. The video record could be used to assess the interpersonal interactions of tellers.
3. The senior author of this report has observed gaming situations where the participants, knowing that they were under television observation still managed after an hour or so to almost completely forget the presence of the cameras.

It would seem that for those work situations where live observer-evaluators are allowed to be present, an audio-video recording system might also be present.

References

Schumacher, S. P., Rudov, M. H. and Valverde, H. H., Evaluation of a Low Cost In-Flight Audio/Video Recording System for Pilot Training, Air Force Human Resources Lab., Technical Report 69-31, 1970.

Description

The interview can be viewed as a particular kind of social interaction situation which lends itself to systematic observation. Some systematic observational systems, of which several were just described, might be used to record interaction during an interview. However, because of its importance, it will be considered here as a separate topic for discussion.

Many of the effectiveness studies that have been done regarding interviewing deal with the counselor-counselee or patient-therapist (psychologist, psychiatrist) interview. Little research has been done on assessing the effectiveness of the physician in the patient-physician clinical interview.

However, a number of studies have been conducted to test the effectiveness of various kinds of interviewing techniques. These "effective" techniques taken as a group, might serve as a basis for developing special behavioral categories which could be used for assessing physician performance in the clinical interview. Several of the interviewing techniques for which there is research support in the literature include the following:

1. Brief expressions of understanding and interest (positive reinforcement) of respondent behavior.
2. Brief expectant pauses.
3. Neutral requests for additional information.
4. Echo or near repetition of the respondent's words.

5. Summarizing or reflection of the respondent's verbal expressions.

There are other basic interviewing techniques that experts agree are effective, such as requesting specific kinds of information; requesting clarification; repeating a primary question.

Finally, characteristics of the interviewer such as facial and vocal expressions as well as conditions of the interview situation (degree of privacy, number of interruptions, etc.) appear to be factors in conducting the "successful" interview.

Judgments About Technique

1. Reliability of the Method. Most studies have shown interviewing to be a rather unreliable means of gathering information, but no data are available on the reliability of assessing the interview situation itself.
2. Validity. Invalid in terms of information gathered. One study showed that, compared with their medical records, respondents under-report illness but at different rates.
3. Current State of Refineness. Needs refinement. Special observation system probably required to assess clinical interview.
4. Feasibility of Using the Method
   a. Cost. $200-400 for each observer per day. If actor-patients are used, cost could be doubled.
   b. People Requirements. One observer and possibly one actor-patient for each clinical interview.

c. Test Environment Requirements. Natural or simulated setting preferred.

d. Suitability for Nationwide Testing Program. Depends on availability of trained observers and actor-patients if required.

5. Type of Behavioral Process Typically Assessed. Could be used to assess interviewing effectiveness of physician based on systematic observation and possibly patient ratings.

6. Appropriateness of NBME R&D Support

a. Refinement Potential. Depends on agreement of experts on categories related to clinical interviewing effectiveness.

b. Adaptability to Assessment of Physicians. Depends on establishment of criteria.

c. Potential for Assessing Non-Cognitive Processes. Good. Should be able to develop both affective and procedural categories once their importance is agreed upon.

References:

1. Conger, J., The Modification of Interview Behavior by Client Use of Social Reinforcement, Behavior Therapy, Vol. 2(1), 52-61, 1971.

2. Froelich, R. E. and Bishop, F. M., Medical Interviewing, The C. V. Mosby Company, 1972.

3. Handbook of Social Psychology, 2nd Edition. Ed. by Gardner Lindzey and Elliot Oronson, Addison-Wesley Publishing Co., Reading, Mass., 1968, Chapter 15.

Overview

In performance tests, the person being evaluated performs tasks which are relevant to his present or future job. Performance tests can range from simulated exercises to performance of job tasks using actual equipment in a natural setting. Simulation is generally employed as a training methodology where there is too much risk and/or too much cost (in terms of resources, time, or personnel) in using the actual work environment. These seem to be reasonable criteria for determining whether to use a simulated or natural work environment for assessment purposes. Specific testing techniques will be described more fully in the following two sections: Simulation and Natural Setting Techniques. What follows is a discussion of performance testing in general and the factors to consider when developing a performance test.

Performance tests should be constructed on the basis of a behavioral job analysis. This analysis determines what skills are required for task performance, the importance or criticality of each task element and an indication of the standards to employ for determining the correct performance of each task element. Apart from criticality, such factors as the time required to perform a given task, type of equipment required, the ability to present the task in a uniform standard manner, and the ability to evaluate an individual's performance with a high degree of objectivity are considerations that must be made for selecting tasks to test.

In all its varied forms, performance tests are widely used in certain skilled trades (apprenticeship programs) and military training

programs in all Services. Post-training performance tests are infrequently used for assessment. This is due to the generally high costs and administrative difficulties of performance testing. However, at times, inexpensive, easily administered performance tests (i.e., paper and pencil "performance" tests) can be used if the required behaviors on the test reflect the appropriate tasks and jobs being assessed. One example is the "identification" test in which the ability to identify photographs or objects that have been printed in a paper and pencil format may have high validity for certain task requirements. Thus, identification tests have been developed for medical laboratory technicians, plumbers, auto mechanics, etc. These are trades in which the individual must be able to identify certain tools, objects and visual stimuli in order to adequately perform his job.

Scoring or evaluation of performance tests can be based on the measurement of the performance in "process," or upon the output or "product" of that performance, or some combination of both. Evaluating "process" is a time-consuming and expensive procedure. The major consideration should be how much the process score contributes to the overall validity of the performance test. Unless higher validity can be demonstrated, it is questionable that the effort and expense involved in process measuring can be justified. There are fewer problems in "product-based" assessment. For one thing, these products can be judged after testing has been completed.

At times, the "process" is the "product." One of the most popular methods of scoring such a performance test is through a sequential checklist. To construct such a checklist, a task is broken down into

the sequential elements which must be correctly performed if the task is to be completed. Each of these elements is then sequentially listed for scoring by an examiner as he observes the performance of the person being tested on the task.

Judgments About Technique

1. Reliability. Depends upon specific test. Psychometric concepts of reliability may not be applicable with regard to performance tests.
2. Validity. Performance tests, as a whole, have impressive "face validity" and credibility because the tasks one must do so closely resemble the job itself.
3. Current State of Refineness. Performance testing as a whole is increasing in use. Thus, more test types, scoring techniques, etc. should be forthcoming.
4. Feasibility of Use
   a. Cost. Usually very high vis a vis paper and pencil tests.
   b. People Requirements. Depending upon scoring approach used, there is often a one examiner per examinee requirement.
   c. Test Environment Requirements. Often on-the-job or simulated equipment or apparatus is needed.
   d. Suitability for Nationwide Testing Program. Unlikely that sufficient resources could be made available for comprehensive performance testing. However, by making use of item sampling techniques, such testing may become feasible.

5. Types of Behavioral Process Typically Assessed by Method. All directly observable behaviors are usually assessed by means of performance tests.

6. Appropriateness for NBME R&D Support

   a. Refinement Potential. Good. Tests need to be developed which can reflect behaviors considered important for "good" or "successful" performance. Such measures, if refined, could justify the costs associated with most performance tests.

   b. Adaptability to MD Assessment. There is no question of the ability to adapt performance tests to assess physicians in most tasks. The advantages of performance tests are realism, objectivity and face or content validity. Against these advantages are the disadvantages of high cost and administrative problems.

   c. Potential for Assessing Non-Cognitive Processes. Only direct way for measuring motor processes. Perceptual and communication processes can also be assessed by performance tests.

References

See specific references under each category of performance tests.

Description

The word simulation has various meanings. It has been defined as an operating model of a real system in that it reflects the behavior of the real system over time. In the context of training and/or assessment, simulation is usually meant to cover situations which model "real world" processes and environments in which trainees and/or examinees can acquire/display skills and knowledges that are required in the "real world."

Gaming, role playing, in-basket exercises, etc. all are instructional techniques that have long been used to teach decision-making and problem-solving skills, and, to a lesser degree, perceptual and communicative abilities. All these techniques are basically simulations -- that is, situations are created that simulate the problems, constraints, and environment of the "real world" work environment. Each of these techniques will be discussed separately. This section is devoted to a general description of the use of simulation as a methodology for assessing performance.

Simulation has been used to teach business management principles and decision-making, international relations, political decision-making and many other subjects at all educational levels. An example is the use of simulated courtrooms in law schools in which organized role-playing occurs. Simulations have been used to train airplane flight crews, plant managers, and in many other job-specific training situations. At one time, training devices such as the Link Trainer,

used for training military pilots, received much publicity. Simulated environments used to train astronauts have recently been discussed. In each of these simulated situations, the goal has been that of instruction. However, assessment of the trainees to see if they have attained the instructional objectives also occurs within the simulated context.

In recent years, simulated training and assessment activity has continued at a considerable level within the three Services. In the civilian community, simulations are on the increase in the preparation of educational/school personnel; the training of guidance and counseling personnel; and the training of para-medical and professional health personnel. The simulation market is understood to be in excess of $100 million a year (Twelker, 1972).

The primary advantage of simulation as an instructional tool, also as an assessment device, is its focus on the operation of the system that permits immediate feedback to the participant of how the system would respond to his own actions. Aside from the high costs and administrative disadvantages of most performance testing techniques, there are other specific problems associated with simulations. For one, there is a need to train individuals to develop and use simulations. Even more important, though, is a lack of any evaluation of simulations. One reason why developers of simulations do not evaluate their products is that no one has clearly in mind what should be evaluated or what might be evaluated. Some users say that a bad simulation is better than anything they now have, so why bother evaluating? Unfortunately, they cannot show how the simulations are

in fact "better." It is of great importance that potential users of simulations approach them with caution because of this lack of evaluative information.

Judgments About Technique

1. Reliability. With regard to reliability, one should try to establish that the performance in a simulated situation is relatively insensitive to random responses. This kind of reliability testing is similar to that in other assessment areas but it is not the traditional psychometric approach used with paper and pencil testing.
2. Validity. The question of validity in a simulation technique pertains to its degree of "realism." One of the ways of validating a simulated situation is to determine if the behavior within a simulated task can or cannot be differentiated from that in the "real world" task that it was designed to simulate. The more realism in the simulation, the more "face" or "content" validity the simulation contains. Another approach discussed by H. Guetzkow (1972) is that of obtaining a "correspondence rating" of the attributes in a simulation as compared to the "real world" situation. The rating of correspondence is either "much," "some," "little," "none," or "incongruent." This rating system is discussed in the reference.
3. Current State of Refineness. Depends upon specific techniques and tasks assessed.

4. Feasibility of Use

   a. Cost. Depends upon specific device, but as is the case with performance tests, costs are relatively high.

   b. People Requirements. Depends upon specific situation, but usually involves a large examiner/examinee ratio.

   c. Test Environment Requirements. Could require job-relevant equipment, access to computer facilities, or merely a pencil and paper format.

   d. Suitability for Nationwide Testing Program. Only suitable if specific measures can be standardized and easily administered.

5. Type of Behavioral Process Typically Assessed by Method. Depending on specific technique, all processes (mediational, motor, perceptual, affective) have been, or can be, assessed by simulation techniques.

6. Appropriateness for NBME R&D Support

   a. Refinement Potential. Good for some techniques but not for others.

   b. Adaptability to MD Assessment. Certain more expensive techniques appear appropriate for in-training use. Others may be adapted more generally.

   c. Potential for Assessing Non-Cognitive Processes. Good. Certain techniques are highly appropriate for interpersonal communication and perceptual process assessment.

References

1. Guetzkow, Harold, et al., Simulation in Social and Administrative Science, Prentice-Hall, 1972.

2. Twelker, Paul A., Some Reflections on Instructional Simulation and Gaming, Simulation and Games, June 1972, 147-153.

Description

Games are used primarily as a training method as are the other simulation techniques discussed in this report. The essential features of a game are: (1) its basic elements are players or actors, each striving to achieve a goal; (2) it is limited to a small set of players; (3) its rules limit the range and define the nature of actions of the players; (4) its rules establish the basic order, sequence and structure within which the action takes place; (5) it is limited in time as well as extensivity with an end defined by the rules; and (6) its rules constitute a temporary suspension of ordinary life activites. It establishes what one might describe as a minute system of activities in a minute social system.

Gaming is distinguished from role playing by its different purpose. Whereas gaming is generally designed to help the participants learn about the system -- either how it operates or how to be an effective part of it -- role playing as such is oriented toward producing greater empathy on the part of the role player with the real life occupant of the role. A good brief history up to 1968 of simulation gaming is that of Boocock and Schild (1968).

The business game attempts to represent the functioning of an industry, a company, or some unit of the above. The game consists primarily of a set of relationships or rules which are derived from theory or from studies of how the firms or industries operate. The trainee usually plans the game of making decisions about what they

would do if they were in a certain role in that firm or in that industry, and, based on their decisions, they would receive feedback regarding the outcome of these decisions. Business management gaming is in widespread use. Graham and Gray (1969) surveyed the 500 largest U.S. corporations and developed a handbook describing over 200 business games in current use.

In the military, war gaming has been employed for many years. The rules and regulations under which the participants play these games are well specified in advance and the usual purpose of the games is to train the participants in applying and refining their skills.

Games and simulations, like in-basket exercises, role-playing, etc., give the learner the opportunity to practice and apply concepts, skills, procedures, and rules that he has previously acquired. This provides a useful environment for assessing these acquisitions. However, a reluctance to use games as assessment devices was uncovered in the literature survey. This may be due to the major disadvantage of the more formalized simulation games -- expense -- the cost of building and testing the models, the cost of implementing the system, the training of operators, computer programs, etc., and the cost of running the system. The demands of instructors, students, and support personnel are substantial. However, in one reported case, a Harvard Business School game used in 1963 required $1200 of computer running costs for a student body of 650 students. Another disadvantage noted is that the competitive aspect of gaming leads some students (examinees) to emphasize "winning the game" at the expense of using it most effectively to learn or perform upon.

Gaming is described here because the authors believe that due to certain of its instructional advantages, the method can provide a good opportunity for applying and testing recently acquired knowledges and skills. We feel that games could serve as a useful assessment device during training as participants can learn from experience without paying the price of wrong decisions made in real life, time can be compressed, feedback is immediate, discussion is realistic, and motivation is high. A participant can gain insight into his own behavior as he interacts with others and can develop a sensitivity to the perceptions, needs, goals, and behavior of others.

Judgments About Technique

1. Reliability. Unknown.
2. Validity. Unknown.
3. Current State of Refineness. Many well-refined instructional games exist.
4. Feasibility of Use
   a. Cost. Very expensive most of the time. However, low cost games may feasibly be developed.
   b. People Requirements. Depends upon game. Administration can be conducted by instructional staff. No additional personnel may be necessary. However, considerable support staff might be required for complex computer-based games.
   c. Test Environment Requirements. Depends upon game requirements. Could range from an ordinary classroom, to one with access to computer facilities, programmers, etc.

d. Suitability for Nationwide Testing Program. Consider formal gaming unsuitable for nationwide testing program.

e. Number of People Tested Each Year. Unknown.

5. Types of Behavioral Process Typically Assessed by Method. Primarily used for instruction in mediational (cognitive) processes.

6. Appropriateness for NBME R&D Support

a. Refinement Potential. Good potential for developing instructional games that could also be used for assessment.

b. Adaptability to MD Assessment. Suitable for assessments during a physician's professional training.

c. Potential for Assessing Non-Cognitive Processes. Interpersonal skills and perceptual processes could be readily assessed using games. Difficult to assess motor skills.

References

1. Boocock, S. S. and Schild, E. O., Simulation Games in Learning, Page Publications, 1968.

2. Graham, R. S. and Gray, C. E., Business Games Handbook, New York: American Management Association, 1969.

ROLE PLAYING

Description

Role playing is a technique whereby the subject is asked to make believe he is relating to a critical object under well defined circumstances. The manner with which the subject deals with the situation is then coded into categories that represent specific properties - direction or magnitude - of an inferred, underlying attitude.

This method has been used extensively for training purposes in business and industry, as well as in psychiatric and intercultural interactions. Not very much research has been done in the area of assessment although the technique has received limited study in assessment of communication processes in physician-patient relationship.

Role taking is a variation of role playing. It involves the subject's playing the role of a well-defined "other" under a given set of circumstances. This method has been mostly used for training purposes, particularly in training individuals to better understand or have "empathy" for another's point of view or reference group. It has had little or no use for assessment purposes.

Judgments About Technique

1. Reliability. Unknown or inappropriate.
2. Validity. Probably face and content valid.
3. Current State of Refineness. Depends on circumstances of setting.

4. Feasibility of Using Method

   a. Cost. Could be very expensive. Possibly over $500 per day for highly trained role players.

   b. People Requirements. One professional role player for each assessment (actor playing role of a patient might be a physician).

   c. Test Environment Requirements. Natural or simulated setting.

   d. Suitability for Nationwide Testing Program. Better suited for training than nationwide testing.

5. Type of Behavioral Process Typically Assessed by Method.

   a. Role Playing. Effectiveness of communication, cross-cultural, sub-cultural.

   b. Role Taking. Patient understanding and empathy.

6. Appropriateness of NBME R&D Support

   a. Refinement Potential. Good if situation is clearly defined.

   b. Adaptability of Assessment of Physicians. Depends on the establishment of criteria.

   c. Potential for Assessing Non-Cognitive Processes. Specifically designed for that purpose.

   d. Potential Validity. Depends on how real situations can be simulated to obtain acceptable face and content validity.

References

1. Biddle, B. J. and Thomas, E. J. Eds., Role Theory: Concepts and Research, John Wiley & Sons, Inc., New York, 1966.

2. Stewart, E. C., Danielian, J. and Foster, R. J., Simulating Intercultural Communication Through Role-Playing, HumRRO TR 69-7, May, 1969.

3. Wohlking, W. and Weiner, H., Structured and Spontaneous Role Playing: Contrast and Comparison, Training and Development Journal, Jan. 1971, p.8-14.

Description

This technique has been employed for many years primarily in industrial settings for purposes of management training. The in-basket is a receptacle used in most offices for incoming and outgoing documents, mail, etc. Thus, in this exercise, an examinee is given the task of performing as if he were a manager at a certain level job. By varying the kind of material in the in-basket, the examiner can set up the situation to require any specific behavior or performance that he wishes to elicit from the examinee. Two references which should give the reader a good description and an account of the history of this technique are Campbell, et al. (1970) and Lopez (1966).

In-basket procedures have been widely used in assessment programs (Bell Systems Management Study) and have been acclaimed for their relevance for predicting managerial effectiveness. The scoring systems used with the in-basket technique generally vary depending upon the behaviors of concern to the investigator.

The Federal Deposit Insurance Corporation recently employed an in-basket exercise successfully for secretarial training. The skills involved were those akin to management training abilities such as office management, supervision, decision-making, human relations, etc. Such subject areas can readily be built into an in-basket exercise.

A critical question to ask of the in-basket technique or other simulation technique is whether or not the behavior required during the simulation reflects behavior required in the actual work situation.

One of the major advantages of an in-basket test is its high face validity. Since it requires performance of an important aspect of work, the use of this test makes sense to examinees. It is probably because of its high face validity that there have been few attempts to validate this test as a measure of actual on-the-job performance.

The in-basket technique, like the other simulation methods, has been primarily used to assess individuals on mediational and cognitive skills, particularly decision-making or problem-solving skills. However, this technique can also be used to assess perceptual and other abilities.

Variations of the in-basket technique have been employed with more than one individual involved. This seems to have a certain potential for use in interpersonal communication skills assessment.

Judgments About Technique

1. Reliability. In a recent study, Meyer (1970) reported split-half reliability coefficients that ranged from .50 to .95.
2. Validity. In the same study, validity scores were obtained between in-basket test scores and criterion performance ratings. Certain skills were better predicted than others, but the concurrent validity coefficients, although significant, were low (.30-.43).
3. Current State of Refineness. This technique is a well-refined simulation method.
4. Feasibility of Use
   a. Cost. Unreported, but has usually not required too great an expense to produce.

b. People Requirements. As with most performance tests, administration has usually been a one-to-one procedure. However, as situations can be set up ahead of time, several examinees could be tested by a single examiner.

c. Test Environment Requirements. Sufficient space and isolation to perform tasks required by simulated excercises.

d. Suitability for Nationwide Testing Program. Alternative forms of standardized exercises would need to be developed before this method could be employed.

e. Number of People Tested Each Year. Unknown.

5. Types of Behavioral Process Typically Assessed by Method. Mediational (decision-making and problem-solving) skills have traditionally been assessed. May be adapted for assessment of interpersonal communication and perceptual abilities.

6. Appropriateness for NBME R&D Support

a. Refinement Potential. Well-refined technique.

b. Adaptability to MD Assessment. Adaptable for assessment of communication skills in certain simulated patient/physician interactions. Highly adaptable for assessing decision-making.

c. Potential for Assessing Non-Cognitive Processes. Moderate for perceptual and communication processes. Not applicable for assessing motor processes.

References

1. Campbell, J. P. et al., Managerial Behavior Performance and Effectivness, New York: McGraw Hill, 1970.

2. Lopez, S. M., Jr., Evaluating Executive Decision Making: The In-Basket Technique, American Management Association Research Study No. 75, 1966.

3. Meyer, H. H., The Validity of the In-Basket Test as a Measure of Managerial Performance, Personnel Psychology, Vol. 22(3), 1970, 297-307.

Description

Performance tests which take place in the work environment using actual equipment are described in this report as "natural setting techniques" as opposed to the "simulations" previously described. These natural setting techniques vary from brief "work sample" or "mini-task" measurements to longitudinal, complete "field" tests.

The prediction of human performance is highly task specific. Information processing requirements, conditions under which the task is to be performed, including environmental constraints, stress, etc. differ from task to task. Quantitative indices that could be used to predict performance are generally lacking, especially with respect to the combined influence of several characteristics found in complex human performance. By developing "mini-task" situations, there is a possibility of uncovering certain quantitative relationships that could be used in the prediction of complex human performance.

Since the late 1940's suggestions have been made to use miniature learning situations as tests for selecting personnel or trainees. The assumption was that learning measures derived from miniature learning situations can predict achievement in subsequent learning situations. This hypothesis was confirmed by a study of Wardrop (1967), and recent studies at HumRRO (Wagner, et al, 1973). These studies generally show that the more the predictive situation is like the criterion situation, the better the resulting validities will be. As more of the skills required in the prediction test are those required subsequently on the criterion, the more valid the predictor will be.

A "mini-task" program was developed for assessing the potential of "low aptitude personnel" for certain Navy ratings. The logic is based upon the assumption that personnel who can learn a sample of the job requirements in a mini-job training situation will demonstrate the same ability on the job. Currently, HumRRO researchers working with the Army's Combat Arms Training Board (CATB) are developing performance test situations to assess the proficiency of enlisted men in various Combat Arms MOSs. These tests would be presented in naturalistic settings.

Although most attempts at measuring teacher effectiveness have been with the use of classroom observation techniques, and global ratings of effectiveness, there has been a movement to produce "micro-criteria" of effectiveness. For example, there were several studies (Westbury and Bellack, 1971) attempting to investigate teacher effectiveness in "explaining" -- one of the required skills for providing instruction to students. A teacher's "effectiveness in explaining" was defined operationally as the ability to present ideas in such a way that the pupils would be able to respond to questions testing the comprehension of those ideas. The lessons used in the assessment procedure were standardized as were test items developed to probe comprehension on the part of students. A similar example was a study to assess the effectiveness of teachers of reading (McNeil, 1971). Performance tests to assess teachers of reading meant giving teachers identical instructional tasks and a sample of the measure to be administered to pupils after the teaching had occurred.

Although rare, comprehensive assessment programs in a natural setting do occur. One recent example is the TEKTITE II human behavior

program conducted by the Navy as part of the SEALAB project (Helmreich, 1971). Systematic quantified observation of human behavior in a natural setting was the heart of this program. The behavior of aquanauts was monitored 24 hours a day during the 182 days that their underwater habitat was occupied. This program had certain significant advantages over other naturalistic research settings. For example, the habitat provided a relatively stable and constant environment. Also, there was the ability to provide excellent audio and video monitoring. Accordingly, the environment provided an excellent situation to observe the aquanaut's criterion performance and adjustment behavior. Between four and six thousand observations were taken on each of the criterion variables on each acquanaut. This particular technique exemplifies the use of natural settings. It clearly provides evidence that following an adaptation period, examinees will perform "realistically" when under observation.

A similar long-term assessment technique which has been used during training is Comprehensive Achievement Monitoring (CAM). This approach systemically measures student achievement in a longitudinal fashion. It tests achievement on every objective of the course by frequent test administration throughout the course. (See Allen, 1970 for further information on this approach).

Judgments About Technique

1. Reliability. Reliability coefficients reported in studies of micro-criteria of teacher effectiveness ranged from .4 to .6. However, as with most performance tests, the usual

concepts of reliability are not applicable. Inter-rater reliability in the TEKTITE II Program ranged from .89 to .97 depending on the variable.

2. Validity. Validation studies performed with work samples have shown concurrent validities ranging from .42 to .66. Actual on-the-job assessments have a basic face validity as the criterion behaviors are directly measured.
3. Current State of Refineness. With the current trend toward performance testing, these techniques should become more refined.
4. Feasibility of Use
   a. Cost. Very expensive in all but small work sample tests.
   b. People Requirements. Usually require a considerable number of observers, monitors, support staff and administrative personnel.
   c. Test Environment Requirements. Depends upon extensiveness of testing program and comprehensiveness of monitoring requirements.
   d. Suitability for Nationwide Testing Program. Would require considerable research to develop suitable on-the-job assessment devices without interfering with ongoing work. Unless criteria and testing conditions could be standardized or well specified, it is not very likely that a suitable natural setting assessment program could be developed.
   e. Number of People Tested Each Year. Unknown

5. Type of Behavioral Process Typically Assessed by Method. All directly observable behaviors.

6. Appropriateness for NBME R&D Support.

   a. Refinement Potential. Good. Natural setting performance tests for physicians can use a lot of refinement. Such an effort appears worthwhile considering the greater validity that would occur with such tests.

   b. Adaptability to MD Assessment. Can be adapted. Key problem to overcome is adaptation of physician to being observed while on the job. TEKTITE II Project indicates this can be overcome if examinee has enough experience in the situation, and expense is no object.

   c. Potential for Assessing Non-Cognitive Processes. Good. All behavioral processes can be so assessed.

References:

1. Westbury, I. and A.A. Bellack, (Eds) Research into Classroom Practices, Teachers College Press, 1971.

2. Wagner, H., et al, Individualized Course Completion Time Predictions: Development of Instruments and Techniques, HumRRO Technical Report, TR 73-25, Nov 1973.

3. Wardrop, James L. A Learning Test Approach to Predicting Classroom Performance, Washington Univ. Report No. R-TR-13, Sept 1967.

4. McNeil, John D., Performance Tests: Assessing Teachers of Reading, Paper presented at California Educational Research Assoc., April 1971.

5. Allen, D.W., et al, An Introduction to Longitudinal Testing Using Item Sampling Techniques: Comprehensive Achievement Monitoring. Paper presented at AERA, March 1970.

6. Helmreich, Robert, The TEKTITE II Human Behavior Program, University of Texas Technical Report No 14, March 1971.

People and Places to Visit

HumRRO Division No. 4 and Combat Arms Training Board personnel at Fort Benning, Georgia.

Description

Adaptive tests are those in which the test items that are presented to a given individual are based on that person's responses to the previous item or to a set of previous items.

Adaptive testing is also known as branch testing, programmed testing, sequential item testing, response-contingent testing, tailored testing, dynamic testing and individualized testing. All these testing approaches involve varying test item presentation according to the characteristics of the individual being tested. The terms identifying these approaches attempt to convey the idea of individualizing or tailoring the testing strategy to a given individual based on responses he has already made to test items already presented. Two recent thorough reviews of adaptive testing are by Wood (1973) and Weiss and Betz (1973). These publications also contain extensive bibliographies of the recent literature in this area.

Proponents of adaptive testing indicate that testing time could be minimized without sacrificing accuracy by giving an individual only those items which are relevant to his ability. This would also decrease guessing considerably since people guess most when items are too difficult for them. Research on adaptive testing has, in general, substantiated these expectations. The results show a definite advantage for adaptive tests over conventional tests in terms of the number of items to be administered to any individual. They show a reliability higher than do conventional tests of the same length. Adaptive tests also give higher correlations with external criteria, requiring conventional tests to

consist of up to twice the number of items as adaptive tests to achieve the same external validity. In some recent studies, adaptive tests containing 40 items correlated as well as did 190-item tests, thus achieving equivalent validities with almost 80% fewer items.

On the negative side, adaptive paper and pencil tests have been shown to require more time to construct than conventional tests and pose difficult scoring problems (i.e., verification of paths examinees take through the test items). Also, a number of examinees have difficulty following the branching instructions, thus invalidating their own test records. It has also been reported that if 500 items are available for adaptive testing, better measurement will often be obtained by selecting 60 of the most discriminating items and administering them as a conventional test rather than by using all 500 in an adaptive testing procedure.

Computer control of adaptive test administration completely avoids the problems inherent in paper and pencil testing. When the computer controls branching, the decisions are completely out of the examinee's hands. The computer presents the test item, records the response, branches an almost infinite number of ways to the next test item and presents it to the examinee.

Computer administered adaptive tests appear to require about 50% less items than conventional tests. Therefore, shorter testing times are needed to achieve the same amount of information. Computerized tests have been shown to yield higher reliability than equivalent conventional tests, shorter testing time and higher validity.

A recent application of adaptive testing is using this approach in performance measurement. This testing technique is described by Kelley

and Prosin (1968). If the measurement system does not change as a conseuqence of the measurement operation and if the problem or stimulus presentation is established in advance the system is not adaptive but fixed. Instead, if the problem or stimulus sequence or the measuring apparatus itself is varied as a result of the performance of the examinee, the measurement system is adaptive. Oral examinations are almost always adaptive in nature (i.e., the questions that are asked are influenced by responses). Psychophysical techniques which vary the stimulus as a function of subject performance are adaptive, with the experimenter performing the adaptive function. Kelley and Prosin (1968) reported on visual acuity tests which have been constructed as adaptive measures.

Application of adaptive testing to the medical profession has taken the form of Tab tests which have been employed by the NBME. These tests require the examinee to select diagnostic tests, procedures or therapy for a patient upon whom background information was provided (in patient management problems). Real life problems, based upon actual medical records are simulated through a set of carefully sequenced problems. At each step of the treatment the examinee is required to make decisions whether to order a particular drug, perform an operation, etc. Whatever he decides, he immediately learns the result of his action and with the new information proceeds to the next incident on the examination. Such an approach, although used primarily for assessing cognitive and problem-solving skills of physicians could be adapted to assess communications and perceptual skills.

Judgments About Technique

1. Reliability. Reliabilities that were reported generally support the view that adaptive tests are at least as reliable as conventional tests of greater length.
2. Validity. Adaptive tests have been shown to be considerably shorter than conventional tests with little or no loss of validity. In several cases adaptive tests were more valid as measured against external criteria.
3. Current State of Refineness. Although many different adaptive testing strategies have been employed in research over the past 25 years, they are still basically experimental and not highly refined. With the increased use of computers in education, computer-based adaptive testing will lead to further refinements.
4. Feasibility of Use.
   a. Cost. High initial development costs for large test item pool needed for adaptive tests. It is generally believed that computerized testing is so much more expensive than paper and pencil testing that it has to be ruled out completely from consideration. However, if a CAI system is in operation, having an examinee answer at a terminal is not far removed from the cost of administering a long group test. It is competitive with individually administered tests, especially if scoring, analysis and reporting costs are included. Moreover, in the future, computer costs can be expected to fall.

b. People Requirements. With the use of computerized testing the requirements for testers decreases. Requirements for test administrators of paper and pencil adaptive tests are equivalent to those of conventional group paper and pencil tests.

c. Test Environment Requirements. The testing time of adaptive test has been shown to be 50% of the time required by equivalent conventional tests. However, the full utilization of adaptive testing requires a time-shared computer facility.

d. Suitability for Nationwide Testing Program. Suitable only if cost were no object in the initial development of the test items, and if a computer facility were available for testing purposes. In addition, examinees need to know how to take tests on such equipment. It is unlikely that computer-based adaptive testing is currently suitable for nationwide testing. A paper and pencil adaptive test would be suitable if sufficient funds were available for test development.

5. Type of Behavioral Process Typically Assessed by Method. Cognitive and psychomotor processes have been assessed by adaptive tests.

6. Appropriateness for NBME R&D Support

a. Refinement Potential. Good for further refinement of adaptive paper and pencil tests. Research available on adaptive testing shows considerable promise though not overwhelming evidence of the superiority of these methods

Measurement Category IV: BEHAVIORAL RECORD AUDIT

Contents

	Index No.	Page No.
Behavioral Record Audit	30	158
Overview		158
Flight Recorders/Ship's Log		158
Personal Histories/Diaries	31	158
Informal Observation/Anecdotal Records	32	159
Unobtrusive Measures	33	159
References		159

Overview

In medical circles considerable attention has been given to the development of assessment procedures based on the audit of medical records. This provides a good illustration of how most evaluation approaches come about -- they are based on what is practical and use whatever information is available. Other than physicians, few job incumbents are required to maintain detailed records of their behavior and decisions.

As we all know airplanes have flight records which record various in-flight maneuvers and aircraft conditions. However, this information is not used for evaluative purposes, unless an accident occurs. The same holds true for a ship's log. It contains a record of ship headings, distances travelled, and perhaps mentions a few of the major events that occurred aboard ship. However, the information is not used to evaluate any one or a group of ship officers unless an accident occurs. Then it and any other information available may be utilized. Our literature search uncovered no studies on the use of records/logs/flight records for the purpose of human performance assessment.

Personal histories and diaries sometimes are used to help in the conduct of a job analysis. They help one maintain a record of daily activities. Diaries are sometimes used to record the critical actions of subordinates so that at personnel evaluation time a supervisor can refer to his personal record of employee performance. Campbell, et al (1970) briefly discuss the use of diaries, but to our knowledge no

research has been reported recently on their use for human performance evaluative purposes.

The place of informal observations and anecdotal records is briefly discussed in Thorndike & Hagen (1969), but we could find no recent studies in which such observations were used to evaluate human performance.

Our literature search uncovered one study reporting on the use of unobtrusive measures. Palmer and McGuire (1973) reported on the "development of 70 unobtrusive measures of six categories (milieu, patient programs, community involvement, treatment, and patient and staff morale) to describe each treatment ward at a state hospital." It was found that "ward profiles" correlated with such criteria as readmission rate, length of stay, and quality rating of the ward. The authors concluded that the "results suggest that unobtrusive measures can contribute significantly to mental health research." Despite the authors' cautious optimism, it would appear that the NBME should await further development in this area before contributing its resources to the study of unobtrusive measures.

References:

1. Palmer, J. & McGuire, F.L. The Use of Unobtrusive Measures in Mental Health Research. Journal of Consulting and Clinical Psychology, 1973, Vol. 40, No. 3, 431-436.

2. Thorndike, R.L. and Hagen, E., Measurement and Evaluation in Psychology and Education, Chapter 14, (3rd Ed.), New York: John Wiley & Sons, Inc., 1969.

ADDITIONAL REFERENCES

The references already cited were the ones used to prepare the descriptive material for this report. The following references are of recent reports which might be of interest to those who wish to study in-depth a particular assessment method or approach. The references have been listed under the following headings.

	Page
Measurement Category I: SELF-PROVIDED INFORMATION	162
Self-Ratings	162
Confidence Testing	164
Measurement Category II: APPRAISAL BY OTHERS	166
Rating Scales: General Articles	166
Rating Scales: Peer/Student/Client Ratings	167
Performance Evaluation Systems	168
Competency-Based Education & Certification Systems	168
Measurement Category III: BEHAVIORAL MEASUREMENT	169
Behavioral Tests	169
Listening	170
Assessment Centers	170
Systematic Observation	170
Systematic Observation: Statistics/Theory	179
Observational Techniques/Methods of Analysis	179
Interpersonal Communication	183
Audio/Visual Recording Systems	184
Interviewing	184
Performance Testing	186
Simulation	186
Gaming	187
Role-Playing	188
In-Basket Exercises	189
Natural Setting Techniques	189
Adaptive Performance Testing	190
Measurement Category IV: BEHAVIORAL RECORD AUDIT	191
Unobtrusive Measures	191
Assessment for the Purpose of Predicting Performance	192
Certification of Health Care Delivery Personnel	194
Training and Evaluation of Health Care Delivery Personnel	195
Evaluation of Training Programs and Students	197
General Review Articles and Texts	198
General/Theoretical Articles	199
Test Directories	202

MEASUREMENT CATECORY I: SELF-PROVIDED INFORMATION

Self-Ratings

1. Baird, Hugh, "Evaluating Teaching episodes" (Concept Teaching Evaluation Form), Brigham Young University.

2. Bedics, Richard A and Webb, Jeaninne N., "Measuring the Self-Evaluation of Teaching Behavior Through the Use of Video Tape," Florida Taxonomy of Cognitive Behavior, Reciprocal Category System, 1971.

3. Boone, Daniel R., and Prescott, Thomas E., Application of Videotape and Audiotape Self-Confrontation Procedures to Training Clinicians in Speech and Hearing Therapy, Part II. Final Report, Denver University, Colorado Speech and Hearing Center, Sept. 1972.

4. Borg, Walter R. Kallenbach, Warren, Morris, Merva, & Friebel, Allen, "Videotape feedback and microteaching in a teacher training model," Journal of Experimental Education, (Sum.), Vol. 37(4) pp9-16.

5. Brammer, Lawrence M. & Springer, Harry C., "A radical change in courselor education and certification," Personnel & Guidance Journal, 1971 (June), Vol. 49(10, pp803-808.

6. Cooper, James M. and Allen, Dwight W., Microteaching: History and Present Status, ERIC Clearinghouse on Teacher Education, Washington D.C, Feb. 1970.

7. Cooper, Joel, Self-Fulfilling Prophecy in the Classroom: An Attempt to Discover the Processes by Which Expectations Are Communicated. Final Report, Princeton University, New Jersey, Oct., 1971

8. Egelston, Richard L.,"Test Achievement: Expectation and Reality," Paper presented at the Annual Meeting of the American Educational Research Association, New York, New York, February 1971.

9. Hiscox, Suzanne B., and Van Mondfrans, Adrian P., "Feedback Conditions and Type of Teaching Skill in Microteaching," Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, April, 1972.

10. Ilgen, Daniel R., "Satisfaction with performance as a function of the initial level of expected performance and the deviation from expectation," Dissertation Abstracts International, 1970(Aug.), Vol. 31(2-B), pp942-943.

11. Ivey, Allen E., "Micro-Counseling and Attending Behavior: An Approach to Pre-practicum Counselor Training," American Personnel and Guidance Association, Washington, D.C., April, 1968.

12. Karasick, Bernard W., and Jewell, Donald O., "Validation of a Student Leadership Scale by Peer and Teacher Nomination Technique," Georgia State College, Atlanta, 1969.

13. Korman, Abraham K., Antonelli, Marie B., Singer, Robert D., and Fekete, Adele F., "The generality of the characteristics of "competent" people," Journal of Vocational Behavior, 1971 (Apr.), Vol. 1(2), pp201-208.

14. Kraemer, Alfred J., "The Development of Cultural Self-Awareness: Design of a Program of Instruction," George Washington University, Alexandria, Virginia, Human Resources Research Office, August, 1969.

15. Kramer, Howard, "Counselors and consulting: Using values in consultation," Journal of College Student Personnel, 1972 (Nov.), Vol. 13(6), pp.534-537.

16. Langer, Philip, "Minicourse: Theory and Strategy," Paper presented as part of a symposium at the annual meeting of the American Educational Research Association, Los Angeles, California, Feb., 1969.

17. Lynch, Helen C., Handbook for Classroom Videotape Recording, Southeastern Educational Corporation, Inc., Atlanta, Georgia, 1969.

18. Martin, Donald G., A Method of Self-Evaluation for Counselor Education. Final Report, Georgia University, Athens College of Education, February, 1968.

19. McKnight, P. C., "Microteaching in teacher training: A review of research," Research in Education, 1971 (Nov.), No. 6, pp24-38.

20. Meers, Gary D., Effects of Various Feedback Media on Psychomotor Achievement, University Microfilms, Ann Arbor, Michigan, May, 1972.

21. Morrison, Virginia B. and Childs, John, "Strategies for the application of videotape in teacher education," Audiovisual Instruction, 1969 (Mar), Vol.14(3), pp43-48.

22. Ohio Education Association, Teacher Evaluation: Interface on Learning, Columbus, Ohio, 1970.

23. Pancrazio, Sally B. and Johnson, William D., "Comparison of Three Teacher Training Approaches in Nonverbal Behaviors Which Encourage Classroom Interaction," Paper presented at annual meeting, AERA, New York, 1971.

24. Roberson, E. Wayne, Teacher Self-Appraisal Source Book, EPIC Evaluation Center, Tucson, Ariz., 1970.

25. Semmel, Melvyn I, et. al., An Information and Technical Manual for the Computer-Assisted Teacher Training System (CATTS), Indiana University, Bloomington Center for Innovation in Teaching the Handicapped, June, 1972.

26. Wagner, Angelika C., "Changing teaching behavior: A comparison of micro-teaching and cognitive discrimination practice," Dissertation Abstracts International, 1971(Sep.), Vol. 32(3-A).

27. Ward, Phillip M., The Use of the Portable Videotape Recorder In Helping Teachers Self-Evaluate Their Teaching Behavior, Calif. University, Berkeley, January, 1970.

28. Webb, Jeaninne Nelson, The Effects of Training in Analysis of Classroom Behavior on the Self-Evaluation of Teaching Performance. Final Report., Alabama University, Tuscaloosa, Sept., 1970.

Confidence Testing

29. Bates, Marina J., "The effect of partial information set on confidence weighting behavior," Dissertation Abstracts International, 1971(Sept.), Vol. 32(30A), pp1178-1179.

30. Echternacht, Gary, The Use of Confidence Testing in Objective Tests, Educational Testing Service, Princeton, New Jersey, Sept., 1971.

31. ....."Use of confidence testing in objective tests., USAF AFHRL Technical Report, 1971(Jul.), No. 71-32.

32. Echternacht, Gary, et al., User's Handbook for Confidence Testing as a Diagnostic Aid in Technical Training, Educational Testing Service, Princeton, New Jersey, July, 1971.

33. Garvin, Alfred D., "Confidence Weighting," Paper presented at the annual meeting of the American Educaitonal Research Association, Chicago, Illinois, April, 1972.

34. Koehler, Roger A., "A comparison of the validities of conventional choice testing and various confidence marking procedures," Journal of Educational Measurement, 1971(Win.), Vol. 8(4), pp297-303.

35. Pugh, Richard C. & Brunza, J. Jay, "Effects of a confidence weighted scoring system on measures of test reliability and validity," Proceedings of the Annual Convention of the American Psychological Association, 1972, Vol. 7(Pt.1), pp51-52.

36. Rippey, Robert M., Scoring and Analyzing Confidence Tests, Final Report, Illinois University, Chicago, May, 1971.

37. Selig, Ernst S., "Confidence testing comes of age," Training & Development Journal, 1972(Jul.), Vol. 26(7), pp.18-22.

38. Shuford, Emir H., and Messengill, H. Edward, Confidence Testing at the Officer Training School, Lackland Air Force Base: September 1968, Shuford-Massengill Corp., Lexington, Mass., Jan., 1969.

39. Suppes, Joan S., "Confidence estimates on the correctness of constructed and multiple-choice responses," Proceedings of the Annual Convention of the American Psychological Association, 1972, Vol. 7(Pt. 1), pp489.490.

MEASUREMENT CATEGORY II: APPRAISAL BY OTHERS

Rating Scales: General Articles

40. Barron, Bruce A., Hirsch, Jules, and Glucksman, Myron, "The construction and calibration of behavioral rating scales," Behavioral Science, 1970(May), Vol. 15(3), pp220-226.

41. Braucht, G. Nicholas, "Analysis and reduction of components of systematic error in ratings," Multivariate Behavioral Research, 1972(Apr.), Vol. 7(2), pp.203-222.

42. Fleiss, Joseph L., "Measuring nominal scale agreement among many raters," Psychological Bulletin, 1971(Nov.), Vol. 76(5), pp378-382.

43. Githens, William H. & Elster, Richard S., "Development of a man-to-man rating scale for evaluating performance," Proceedings of the 81st Annual Convention of the American Psychological Association, Montreal, Canada, 1973, Vol. 8, pp761-762.

44. Horner, Walter R., et al., Final Pilot Performance Rating Scales, American Institutes for Research in the Behavioral Sciences, Pittsburg, Pennsylvania, August, 1970.

45. Masters, James R., "Reliability as a function of the number of categories of a summated rating scale," Dissertation Abstracts International, 1973(Feb.), Vol. 33(8-A), pp4180-4181.

46. Moores, B., "Variability in concept of standard in the performance rating process," International Journal of Production Research, 1972(Apr.), Vol. 10(2), pp167-173.

47. O'Gorman, J. G., "Rating in performance appraisal: A bibliography," Australian Army Psychological Research Unit Research Report, 1972(Mar.), No. 4/72.

48. Ostrom, Thomas M., "The bogus pipeline: A new ignis fatuus?" Psychological Bulletin, 1973(Apr.), Vol. 79(4), pp252-259.

49. Palef, S. and Stewart, C. P., "Development of a rating scale (CR 1's & 2's)," Studies in Personnel Psychology, 1971(Oct.), Vol. 3(2), pp7-20.

50. Quinn, James L., "Bias in performance appraisal," Personnel Administration, 1969(Jan.), Vol. 32(1), pp40-43.

51. Shepard, Roger N., Romney, A. Kimball and Nerlove, Sara B. (Eds.), Multidimensional scaling: Theory and applications in the behavioral sciences: I. Theory, New York, New York, Seminar Press, 1972.

52. Start, K. B., "Teacher assessment as a problem of sampling within a complex composite criterion," International Review of Applied Psychology, 1972(Apr.), Vol. 21(1), pp53-60.

53. Wicklund, Robert A., Slattum, Valerie, and Solomon, Ellen, "Effects of implied pressure toward commitment on ratings of choice alternatives," Journal of Experimental Social Psychology, 1970(Oct.), Vol. 6(4), pp449-457.

54. Wittenborn, J. R., "Reliability, validity, and objectivity of symptom-rating scales," Journal of Nervous & Mental Disease, 1972 (Feb.), Vol. 154(2), pp79-87.

Rating Scales: Peer/Student/Client Ratings

55. Case, Chester H., and Brown, Steven A., Introducing SPT/VTR:P (Sequenced Peer Teaching/Video Tape Recording: Playback), 1970

56. Levinthal, Charles F., Lansky, Leonard M., and Andrews, O. Ernest, "Student evaluations of teacher behaviors as estimations of real-ideal discrepancies: A critique of teacher rating methods," Journal of Educational Psychology, 1971(Apr.), Vol. 62(2), pp104-109.

57. Mayfield, Eugene C., "Value of peer nominations in predicting life insurance sales performance," Journal of Applied Psychology, 1972, (Aug), Vol. 56(4), pp319-323.

58. McIlvaine, Joseph F., "An investigation of client ratings of effectiveness for beginning counselor trainees," Dissertation Abstracts, 1970(Dec.), Vol. 31(6-A), pp2688-2689.

59. Rodin, Miriam and Rodin, Burton, "Student evaluations of teachers," Science, 1972(Sep.), Vol. 177(4055), pp1164-1166.

60. Schein, Virginia E., "The relationship between accuracy of subordinate's perception of his supervisor and job performance," Dissertation Abstracts International, 1970(Sep.), Vol. 31(3-B), pp1581-1582.

61. Smith, Alden W., Students Evaluate Teaching, 1969.

62. Stemnock, Suzanne K., The Evaluation of Noncertificated Personnel, Educational Research Service Circular Number Four, American Association of School Administrators, Washington, D.C., June, 1969.

63. ....., Evaluating Teaching Performance, Educational Research Service Circular Number Three, American Association of School Administrators, May, 1969.

64. Touq, M. S., et al.,"Criterion-Referenced Validity of Student Ratings of Instructors," Paper presented at annual meeting of American Educational Research Association, New Orleans, Louisiana, February 25-March 1, 1973).

Performance Evaluation Systems

65. Hane, Edward Z., "Performance evaluation based on multidimensional job behaviors," Dissertation Abstracts International, 1971(May), Vol. 31(11-B), p6966.

66. Lipsett, Laurence, "Evaluating the performance of service agency counselors," Vocational Guidance Quarterly, 1971(Mar.), Vol. 19(3), pp206-210.

67. Whitley, Richard and Frost, Penelope A., "The measurement of performance in research," Human Relations, 1971(Apr.), Vol. 24(2), pp161-178.

Competency-Based Education & Certification Systems

68. Horan, John J., "Behavioral goals in systematic counselor education," Counselor Education & Supervision, 1972(Mar.), Vol. 11(3), pp162-170.

69. Koenigsberg, Lewis A. and Reilly, Robert R., An Investigation of the Reliability and Validity of Selected Occupational Competency Examinati ns and Their Use in Evaluating Prospective Trade and Industrial Teachers. Final Report., New York State Education Dept., Albany, New York, June, 1968.

MEASUREMENT CATEGORY III: BEHAVIORAL MEASUREMENT

Behavioral Tests

70. Barton, Proger B., "JSE test reviews: Introduction and philosophy," Journal of Special Education, 1970(Win.), Vol. 4(1), pp115-116.

71. Briggs, Peter F. and Tellegen, Auke, "Development of the Manual Accuracy and Speed Test (MAST)," Perceptual & Motor Skills, 1971 (Jun.), Vol. 32(3), pp923-943.

72. Cappon, Daniel and Banks, Robin,"'The birth of a test': The Sensory Quotient," Perceptual & Motor Skills, 1970(Oct.), Vol. 31(2), pp408-410.

73. Ellison, John W., "Visual Tests: An Alternative," Educational Technology, May 1973, Vol. 13, pp50-52.

74. Farina, Alfred J. and Wheaton, George R., Development of a Taxonomy of Human Performance: The Task Characteristics Approach to Performance Prediction, American Institutes for Research Technical Report 7, February, 1971.

75. Gross, Lynne S., A Case for Visual Testing, Acolyte Publications, Los Angeles, California, April, 1969.

76. Jones, Marshall B., Individual Differences in Motor and Verbal Skills. Final Report, Tulane University, Louisiana, New Orleans, August, 1968.

77. Musgrove, Dolores M., "A factor analytic study of perceptual motor attributes as measured by selected test batteries," Dissertation Abstracts International, 1971(Oct), Vol. 32(4-A), pp1904-1905.

78. Norton, Robert Ellsworth, Using Programmed Instruction With and Without Self-Instructional Practice to Teach Psychomotor Skills. Final Report., Cornell University, Ithaca, New York, June 1967.

79. Sanders, James H., Valentine, Lonnie D., and McGrevy, David F., "The development of equipment for psychomotor assessment," USAF AFHRL Technical Report, 1971(Jul.), No. 71-40.

80. Theologus, George C., and Fleishman, Edwin A., Development of a Taxonomy of Human Performance: Validation Study of Ability Scales for Classifying Human Tasks, American Institutes for Research Technical Report 10, April 1971.

81. Zullo, Thomas G., "Patterns of Perceptual Motor Skills in First and Fourth Year Dental Students," Paper presented at the annual meeting of the American Educational Research Association, New York, February, 1971.

Listening

82. Duker, S., "Teaching listening: Recently developed programs and materials," Training & Development Journal, 1970(May), Vol. 24(5), pp11-15.

Assessment Centers

83. Bray, Douglas W., and Moses, Joseph L., "Personnel Selection," Annual Review of Psychology, 1972, Vol. 23, pp545-576.

84. Moses, Joseph L., "Assessment center performance and management progress," Studies in Personnel Psychology, 1972(Spr.), Vol. 4(1), pp7-12.

Systematic Observation

85. Abramson, Theodore, Development of Improved Techniques for Establishing the Reliability of Observation Ratings. Final Report, City University of New York, January, 1970.

86. .....Reliability of an Observation of Teachers' Classroom Behavior, School University Teacher Education Center, New York, 1969.

87. Amidon, Edmund,"Interaction Analysis and Microteaching Skill Development in Teaching," Paper preserted. American Psychological Association, Washington, D.C., 1969.

88. Amidon, Peggy, Nonverbal Interaction Analysis. A Method of Systematically Observing and Recording Non-verbal Behavior, Association for Productive Teaching, Minneapolis, Minnesota, 1971.

89. Beegle, Charles W., and Brandt, Richard M. (Eds.), Observational Methods in the Classroom, Association for Supervision and Curriculum Development, Washington, D.C., 1973.

90. Bellack, Arno A., Methods for Observing Classroom Behavior of Teachers and Students., Presented at a conference sponsored by Padagogisches Zentrum, Berlin, November, 1968.

91. Belland, John, et al.,"Analyzing Teacher Questions: A Comparative Evaluation of Two Observation Systems," Paper presented at AERA annual meeting, New York, 1971.

92. Bemis, Katherine A., and Liberty, Paul G., "Southwestern Cooperative Educational Laboratory Interaction Observation Schedule (SCIOS): A System for Analyzing Teacher-Pupil Interaction in the Affective Domain," Paper presented at the annual convention of the American Educational Research Association, Minneapolis, Minnesota, March, 1970.

93. Bersoff, Donald N. and Moyer, Dale, "Positive Reinforcement Observation Schedule (PROS): Development and applications to educational settings," Proceedings of the 81st Annual Convention of the American Psychological Association, Montreal, Canada, 1973, Vol. 8, pp713-714.

94. Bosch, Albert C., Relationships of Teaching Patterns to Indices of Classroom Verbal Interaction Behavior: A Further Analysis and Synthesis of Classroom Verbal Interaction Data Using Descriptive Indices of Behavior and Teaching Pattern Analysis, University Microfilms, 1972.

95. Blumberg, Arthur and Cusick, Philip, "Supervisor-Teacher Interaction: An Analysis of Verbal Behavior," Paper presented at annual meeting, AERA, Minneapolis, Minnesota, 1970.

96. Bohlken, Robert and Giffin, Kim, "A Paradigm for Determining High School Teacher Effectiveness," Paper presented at the Annual Meeting of the Speech Communication Association, New Orleans, Louisiana, December, 1970.

97. Boyd, Robert E., "Counseling Uses of the Hill Interaction Matrix," Paper presented at the Annual Convention of the American Personnel and Guidance Association, Atlantic City, New Jersey, April, 1971.

98. Brown, Bob Burton, An Investigation of Observer-Judge Ratings of Teacher Competence. Final Report, Florida University, Gainesville, January, 1969.

99. Brown, Bob Burton and Bane, Robert King, "Multidimensionality: A Technique for Studying the Classroom," Paper presented at annual meeting, AERA, Minneapolis, Minnesota, 1970.

100. Brown, Bob Burton, et al., Systematic Observations: Relating Theory and Practice in the Classroom, Florida University, Gainesville, Institute for Development of Human Resources, February, 1969.

101. Brunner, Ellen, "Profile of Interaction in the Classroom. A Quick Feedback of Interaction Analysis," Paper presented at AERA convention, Minneapolis, Minnesota, 1970.

102. Buzzotta, V. R. and Lefton, R. E., "A systematic approach to interpersonal behavior," Training & Development Journal, 1971 (Feb.), Vol. 25(2), pp8-14.

103. Campbell, James Reed, "Pattern Analysis - A Macroscopic Development for Interaction Analysis," Paper presented at the annual meeting of the National Association for Research in Science Teaching (46th, Detroit, Michigan, March 1973).

104. Coates, Carolie, et al., "Observation Scales of Classroom Atmosphere and Student Behavior: A Replication and Refinement," Paper presented at the annual meeting of the American Educational Research Association, Minneapolis, Minnesota, March, 1970.

105. Cobb, Joseph A. and Hops, Hyman, Coding Manual for Subject/Peer/Teacher Sequential Interactions in Academic Survival Skill Settings. Robert No. 4., Oregon University, Department of Special Education, Eugene, Oregon, September, 1971.

106. Davis, O. L., Jr. et al., Studying the Cognitive Emphases of Teachers' Classroom Questions, Texas University, Research and Development Center for Teacher Education, Austin, Texas, April, 1969.

107. Dempsey, Richard A., and Breyer, Norman L., Staff Development and Evaluation, A. C. Croft, Inc., Santa Monica, California,1971.

108. Dennis, Virginia C. and Powell, Evan R., "Nonverbal Communication in Across-Race Dyads," Paper presented at the 80th Annual Convention of the American Psychological Association, Honolulu, Hawaii, September, 1972.

109. Diamond, Michael I, "A comparison of three content analytic approaches to the analysis of interviewer verbal behavior in the initial psychiatric interview," Dissertation Abstracts International, 1970(Oct.), Vol. 31(4-B), pp2301-2302.

110. Dupre, Robert W., The Effects of Counselor Response Upon the Premature Termination of Clients, Louisiana Polytechnic Institute, Ruston, 1970.

111. Duetscher, John, The Study of Non-Verbal Action of Counselees. Final Report. WSU-cord, Wisconsin State Universities Consortium of Research Development, Stevens Point, July, 1969.

112. Educational and Cultural Center, Behavioral Analysis Instrument for Teachers, Syracuse, New York.

113. Emmer, Edmund T., "Classroom Observation Scales," Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, Louisiana, February, 1973.

114. Emmer, Edmund T. and Peck, Robert F., "Dimensions of Classroom Behavior," Paper presented at the Annual Meeting of the American Educational Research Association, New York, New York, February, 1971.

115. Evans, Thomas P., "Flanders System of Interaction Analysis and Science Teacher Effectiveness," Paper presented at the annual NARST meeting, Minneapolis, Minnesota, March, 1970.

116. Fink, Albert H., Fink Interaction Analysis System," Paper presented at the Annual Meeting of the American Education Research Association, New York, New York, February, 1971.

117. Flynn, John M. et al., An Analysis of the Role of the Teacher in an Innovative Prototype School, Part II, Appendices. Final Report, Nova University, Ft. Lauderdale, Florida, February, 1969.

118. Forbes, Joan, "Micro-PIC. A Simple Form of the Profile of Interaction in the Classroom," Paper presented at AERA convention, Minneapolis, Minnesota, March, 1970.

119. French, Russell L., "A Study of Communication Events and Teacher Behavior: Verbal and Nonverbal," Paper presented at the annual meeting, AERA, Minneapolis, Minnesota, March 1970.

120. French, Russell and Galloway, Charles M., Communication Events: A New Look at Classroom Interactions, 1968.

121. Furst, Norma, et al., Interaction Analysis: Selected Papers, Association of Teacher Educators, Washington D.C., November, 1971.

122. Geisinger, Robert W., Interaction Analysis of Value-Clarification Behaviors, Pennsylvania State Department of Education, Harrisburg, October, 1970.

123. Gibson, Dennis L., "The HIM-G as a Rapid Method for Verbal Interaction Analysis in a Small Group," Paper presented at the American Personnel and Guidance Association convention, New Orleans, Louisiana, March, 1970.

124. Grant, Barbara M., "Teach Non-Verbal Activity," Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Illinois, April, 1972.

125. Grimm, Jeffrey A., "A technique for minimizing subject-observer looking interactions in field settings," Journal of Experimental Child Psychology, 1972(Dec.), Vol. 14(3), pp500-505.

126. Hall, Gene E., "The Use of Interaction Analysis in the Modification of Science Teacher Behavior," Presented at the National Association for Research in Science Teaching Meeting, Minneapolis, Minnesota, March, 1970.

127. Heger, Herbert K., "Verbal and Nonverbal Classroom Communication: The Development of an Observational Instrument," Paper presented at annual meeting, AERA, Minneapolis, Minnesota, March, 1970.

128. ....., Analyzing Verbal and Nonverbal Classroom Communications, 1968.

129. Hill, John C., "Change of Content Development Patterns Observed in Classroom Communication Behaviors Due to In-Service Training in Content Strategies," Paper presented at annual meeting, AERA, New York, 1971.

130. Hoehn, Lilburn P., Ed., Teaching Behavior Improvement Program, Michigan-Ohio Regional Education Labs., Inc., Detroit, July, 1969.

131. Hough, John B., "Ideas for the Development of Programs Relating to Interaction Analysis," Title III PACE, Projects to Advance Creativity in Education, Conference report, Michigan State Board of Education, Lansing, Michigan, August, 1966.

132. Jones, Margaret Hubbard, Reliability of Coding of the System for the Analysis of Classroom Communication (SACC). Center for the Study of Evaluation Working Paper Number Two, California Univ., Center for the Study of Evaluation, Los Angeles, California, November, 1969.

133. Jorgensen, Gary Q., et al., Interpersonal Relationships. Factors in Job Placement, Utah University, Regional Rehabilitation Research Institute, Salt Lake City, March, 1968.

134. Kiser, Luther, et al., Project Evaluation: The Effect of Selected Media Feedback upon the Interactive Behavior of Student Teachers. Final Report, Missouri University, Columbia College of Education, October, 1969.

135. Komulainen, Erkki, Investigations into the Instructional Process. VIII. On the Problems of Variable Construction from Flanders' Interaction Matrix with Special Emphasis on the Stochastic Nature of Classroom Communication, Helsinki University, Institute of Education, Finland, February, 1973.

136. ....., Investigations into the Instructional Process II. Objectivity of Coding in a Modified Flanders Interaction Analysis, Helsinki University, Institute of Education, Finland, December, 1970.

137. Korth, Willard W., et al., A Study of the Relationships Between Measures of Teacher-Pupil Verbal Interaction and Student Assessment of Classroom Practices, Pittsburgh University, Pennsylvania, 1971.

138. Ladd, George Thomas, Determining the Level of Inquiry in Teachers' Questions, University Microfilms, Ann Arbor, Michigan, 1969.

139. Lamb, Wesley A., A proposal for evaluation of the reliability-agreement of clinical observation procedures, Proceedings of the 81st Annual Convention of the American Psychological Association, Montreal, Canada, 1973, Vol. 8, pp527-528.

140. Langer, Philip, and Allen, G. Edward, "The Minicourse as a Tool for Training Teachers in Interaction Analysis," Paper presented at the annual meeting of American Educational Research Association, Minneapolis, Minnesota, March, 1970.

141. Lawson, Dene R., "Indicators of Teacher Ability to Relate to Students," Paper presented at AERA annual meeting, New York, 1971.

142. Limbacher, Philip C., and Rosenshine, Barak, "Relationship of High-Inference and Low-Inference Observation Measures," Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Illinois, April, 1972.

143. Long, John V., The Effect Upon Verbal Interaction Data of the Medium Employed in the Codification Process, Rhode Island Univ., Kingston. Curriculum Research and Development Center., 1971.

144. Luft, Max, and Bemis, Katherine A., Video Tape Techniques for Establishing Inter-Rater Reliability, Southwestern Cooperative Educational Lab., Albuquerque, New Mexico, March, 1970.

145. McGaw, Barry, Wardrop, James L., and Bunda, Mary A., "Classroom observation schemes: Where are the errors?" American Educational Research Journal, 1972(Win.), Vol. 9(1), pp13-27.

146. Medley, Donald M., "Supervisor Use of Observation Systems," Paper presented at Supervision of Instruction Symposium 2: Observation Systems and the Supervisor, January, 1972.

147. ....., Studying Teacher Behavior with the OScAR Technique, Proceedings: Conference on Instructional Methods and Teacher Behavior, Berkeley, California, November, 1966, pp47-50.

148. Mark, Gordon M. A., A Taxonomy-Based Verbal Performance Instrument

149. Morse, Kevin R., and Davis, O. L., Jr., The Questioning Strategies Observation System (QSOS), Texas University, Research and Development Center for Teacher Education, Austin, May, 1970.

150. Moskowitz, Gertrude, Interaction Analysis Gives Insight into Student-Teacher Exchanges, Georgia State University, Atlanta, 1972.

151. Park, James, "The Use and Development of Videotape Tests to Assess Achievement of Educational Objectives: Application in Educational Psychology," Paper presented to the CAPE Conference, Montreal, 1972.

152. Ober, Richard L., et al., "Simultaneous Use of Four Different Observational Systems To Assess Student Teacher Classroom Behavior," Paper presented at the annual meeting of AERA, Minneapolis, Minn., March, 1970.

153. Ober, Richard L., The Reciprocal Category System (RCA): An Observational System Designed To Assess Teacher-Student Classroom Verbal Interaction, Richard L. Ober, College of Education, Univ. of South Florida, Tampa, Florida, 1968.

154. Olivier, Kenneth S., Jr., Data Collection and Reduction Procedures Used for the System for the Analysis of Classroom Communications (SACC). Center for the Study of Evaluation Working Paper No. 5., California University, Center for the Study of Evaluation, Los Angeles, 1969.

155. Ramirez, Judith Valla, Teacher Behavior in Role-Playing: A Study in Interaction Analysis. Research and Development Memorandum No. 43., Stanford University, Stanford Center for Research and Development in Teaching, California, February, 1969.

156. Reynolds, William W., Jr., et al., "The Classroom Observational Record," Paper presented at the Annual Meeting of the American Educational Research Association, New York, New York, February, 1971.

157. Rose, Harold, "The Effect of Sociability on the Type and Rate of Interaction in Small Groups," Paper presented at the National Seminar on Adult Education Research, Toronto, February, 1969.

158. Rose, Norman Harold, The Relationship between Sociability and Type and Rate of Interaction, University Microfilms, Ann Arbor, Michigan, 1969.

159. Rosenshine, Barak, et al., Empirical Validation of a New Category System: One Example, Temple University, Philadelphia, Pennsylvania, 1971.

160. Rousseau, Joseph F., "A Study of Verbal Behaviors Engaged in by College Student Teaching Supervisors in Conferences with Cooperating Teachers and Student Teachers," Paper presented at the Annual Meeting of the American Educational Research Association, April, 1972.

161. Samph, Thomas, Observer Effects on Teacher Behavior, 1968.

162. Schutte, Alfred John, An Exploratory Investigation of Micro-Teaching as a Pre-Service Technique and Its Evaluation, Using the Flanders System of Interaction Analysis, University Microfilms, Ann Arbor, Michigan, 1971.

163. Sharpe, Donald M., Isolating Relevant Variables in Student Teaching Assessment. Final Report, Indiana State University, School of Education, Terre Haute, January, 1969.

164. Smidchens, Uldis, and Roth, Rod, Michigan-Ohio Regional Educational Laboratory's Computerization of Interaction Analysis Project, Michigan-Ohio Regional Educational Lab., Inc., Detroit, July, 1968.

165. Southwell, Reba K., and Webb, Jeaninne N., "Development and Validation of an Observation System for Analyzing Teaching Roles," Paper presented at the AERA, Chicago, Illinois, April, 1972.

166. Spaulding, Robert L., "Observational Methodology in the Experimental School Setting," Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, California, February, 1969.

167. Sprague, Nancy Freitag, "Inquiry Dialogue in the Classroom," Paper presented at the Annual Convention, American Educational Research Association, New York, New York, February, 1971.

168. Steen, Margaret T., and Lipe, Dewey, "The Use of the Teacher Observation Scale in the Development of the PLAN Teacher Training Program," Paper read at the annual meeting of the American Educational Research Association, Minneapolis, Minn., March, 1970.

169. Stukat, Karl-Gustaf, Microteaching, National Swedish Board of Education, Stockholm, November, 1972.

170. Swanson, Maria Antonieta Medina, Interaction Analysis in Foreign Language Teaching: A Rationale, University of Texas at Austin, May, 1971.

171. Turney,Mildred I, A Comparison of Varied Time Periods of Microteaching in the Development of Interpersonal Relationships in Teaching. Final Report, Wisconsin State Universities Consortium of Research Development, Stevens Point, December, 1969.

172. Twelker, Paul A., Interaction Analysis and Classroom Simulation As Adjunct Instruction in Teacher Education. Final Report, Oregon State System of Higher Education, Teaching Research Div., Monmouth, February, 1968.

173. Victoria, J. James, An Investigation of Nonverbal Behavior of Student Teachers. Final Report, Pennsylvania State University, University Park, Pennsylvania, April, 1970.

174. Walker, Rob, and Adelman, Clem, Time-Lapse Photography in Recording Classroom Events, Chelsea College of Science and Technology, London, England, January, 1971.

175. Webb, Janinne, et al., "Observation as a Methodology," Paper presented at the American Educational Research Association meeting, Minneapolis, Minnesota, March, 1970.

176. Webb, Jeaninne Nelson, and Brown, Bob Burton, "Establishing Reliability and Validity Estimates for Systematic Classroom Observation," Presented at the 1969 American Educational Research Association meeting, Los Angeles, California, February, 1969.

177. Whelan, Sister Mariellen, An Exploratory Study of Interaction Analysis in the College Classroom, 1972.

178. Wood, Margaret W., et al., "Training Student Teachers Using the Reciprocal Category System of Interaction Analysis," Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, California, February, 1969.

179. Wood, Samuel E., et al., Collecting and Analyzing Data Yielded by the Multidimensional Technique for Observing Classroom Behavior, 1969.

180. Wood, Samuel E., et al., "A Factor Analysis of Three Sets of Simultaneous Collected Observational Data: Theory and Implications," Paper read at Symposium of American Educational Research Association program, Los Angeles, California, February, 1969.

181. Wulff, John W., Analysis of Verbal Interaction in Supervisory Conferences with Student Teachers, 1971.

182. Zinser, Robert W., Communicating Classroom Teacher Behavior. A Systematic Approach, Cleveland Public Schools, Ohio, 1971.

Systematic Observation: Statistics/Theory

183. Baker, J. Philip, Using Generalizability Theory and Multifacet Designs in the Validation of a Classroom Observation Instrument, Stanford University, Stanford Center for Research and Development in Teaching, California, September, 1971.

184. Conklin R. C. and Ogston, D. G., "The effect of varying the mode of information on interpersonal judging accuracy," Western Psychologist, 1970(May), Vol. 1(3), pp100-105.

185. Gregory, Thomas B., A Computer Program for Calculation of Scott's Coefficient of Observer Reliability, Texas University, Research and Development Center for Teacher Education, Austin, Sept., 1969.

186. Maxwell, A., "Agreement among raters," British Journal of Psychiatry, 1971(June), Vol. 118(547), pp659-662.

187. Pesaresi, Edward, "The effect of observation by experimenters differing in authority on gross motor performance," Dissertation Abstracts International, 1970(Sep.), Vol. 31(3-A), p1374.

Observational Techniques/Methods of Analysis

188. Abramson, Theodore, and Spilman, Helen, Observation Instruments and Methodology and Their Application in The Classroom: An Annotated Bibliography, City University of New York, New York Office of Teacher Education, September, 1971.

189. Afandor, Josef C., "Interactions analysis as a measure of counselor effectiveness, Dissertation Abstracts International, 1972(Jan.), Vol. 32(7-A), p3678.

190. Amidon, Edmund, "Interaction Analysis and Supervision," Paper presented at Supervision of Instruction Symposium 2: Observation Systems and the Supervisor, January, 1972.

191. Anderson, Ronald D., et al., "Development of a Verbal and Non-Verbal Observation Instrument," Paper presented at the Annual Meeting of the American Educational Research Association, Minn., Minnesota, March, 1970.

192. Berk, Richard A., and Adams, Joseph M., "Establishing rapport with deviant groups," Social Problems, 1970(Sum.), Vol. 18(1), pp102-117.

193. Bersoff, Donald N., Ericson, Colleen R., "A Precise and Valid Measure of Behavior and Behavior Change in the Classroom," Paper presented at the 79th Annual Convention of the American Psychological Association, Washington, D.C., September, 1971.

194. Blumberg, Richard W., "Training clinical specialists for behavioral observation," Psychological Reports, 1971(April), Vol. 28(2), pp568-570.

195. Brophy, Jere E., and Good, Thomas L., Teacher-Child Dyadic Interaction: A Manual for Coding Classroom Behavior. Report Series No. 27, Texas University, Research and Development Center for Teacher Education, Austin, 1969.

196. Burton, Roger V., "An inexpensive and portable means for one-way observation," Child Development, 1971(Sep.), Vol. 42(3), pp959-962.

197. Byers, Burton H., "Testing Interpersonal Communciation Performance," Paper presented at the Annual Meeting of the International Communication Association, Montreal, April, 1973.

198. Chandler, David R., "Construct validity of a scale of accurate empathy in psychotherapists," Dissertation Abstracts International, 1971(Sept.), Vol. 32(3-B), pp1836-1837.

199. Dunning, Gail B., "The identification and measurement of nonverbal communication in the counseling interview," Dissertation Abstracts International, 1972(May), Vol. 32(11-A), pp6124-6125.

200. Engel, George L., "The education of the physician for clinical observation: The role of the psychosomatic (liaison) teacher," Journal of Nervous & Mental Disease, 1972(March), Vol. 154(3), pp159-164.

201. Galloway, Charles M., "The Nonverbal: An Approach for Supervisors," Paper presented at Supervision of Instruction Symposium 2: Observation Systems and the Supervisor, January, 1972.

202. .....Teaching is Communicating: Nonverbal Language in the Classroom, Association for Student Teaching, Washington, D.C., 1970.

203. .....Nonverbal Communication: A Needed Focus, 1968.

204. Geertsma, Robert H., "Observational methods," Clinical assessment in counseling & psychotherapy, New York, New York, Appleton-Century-Crofts, 1972.

205. Good, Thomas L., and Brophy, Jere E., Analyzing Classroom Interaction: A More Powerful Alternative, Texas University, Research and Development Center for Teacher Education, Austin, 1969.

206. Green, Paul F., "A participant-observer evaluation of an inservice teacher workshop," Dissertation Abstracts International, 1970 (Oct.), Vol. 31(4-A), p1536.

207. Hansen, John H. and Anderson, Robert A., Trainer's Manual: Interaction Analysis, Oregon University, Portland, December, 1969.

208. Harmatz, Morton G., "Observational study of ward staff behavior," Exceptional Children, 1973(Apr.), Vol. 39(7), pp554-558.

209. Harris, Alice M., "Observer effect on family interaction," Dissertation Abstracts International, 1970(Sep.), Vol. 31(3-A), p1075.

210. Heagarty, Margaret C., and Robertson, Leon S., "Slave doctors and free doctors: A participant observer study of the physician-patient relation in a low-income comprehensive-care program," New England Journal of Medicine, 1971(Mar.), Vol. 284(12), pp636-641.

211. Hinde, R. A., Ed., Non-Verbal Communication, Cambridge University, England, 1972.

212. Hunt, David E., "Adaptability in interpersonal communication among training agents," Merrill-Palmer Quarterly, 1970(Oct.), Vol. 16(4), pp325-344.

213. Hunt, H. Thomas, and Anapol, Malthon M., "Studying Trial Variables through Videotaped Trials," Paper presented at the Annual Meeting of the Speech Communication Association, Chicago, December, 1972.

214. Janzen, Frederick V., "Client-counselor perceptual congruency as measured by the interaction scale," Dissertation Abstracts International, 1970(Nov.), Vol. 31(5-A), pp2183-2184.

215. Masonis, Edward J., "Problems and Procedures in Planning a Situation Based Video Test on Teaching," Paper presented at the Annual Meeting of the American Personnel and Guidance Association, Atlantic City, New Jersey, April, 1971.

216. McGinnis, Ruth L, "A psychometric evaluation of Carkhuff's Discrimination Test for counselor effectiveness," Dissertation Abstracts International, 1972(Feb.), Vol. 32(8-B), p4847.

217. Mitchell, Kevin M., "An ecological orientation to the helping process," Rehabilitation Research & Practice Review, 1970(Sum.), Vol. 1(3), pp9-17.

218. Nuthall, G. A., "A Review of Some Selected Recent Studies of Classroom Interaction and Teaching Behavior," in Gallagher, James J., Nuthall, Graham A., and Rosenshine, Barak, Classroom Observation, AERA Monograph in Curriculum Evaluation, No. 6, Rand McNally and Company, Chicago, 1970, pp6-29.

219. O'Connor, J. Regis, "Kinesics, Communication and Group Interaction," Paper presented at the Convention of the Southern Speech Communication Assn., Bowling Green, Kentucky, April,1972.

220. Roberson, E. Wayne, Developing Observation Systems, Educational Innovators Press, Tucson, Arizona, 1970.

221. Rosenshine, Barak, "Evaluation of classroom instruction," Review of Educational Research, 1970(Apr.), Vol. 40(2), pp279-300.

222. Sandefur, J. T., and Bressler, Alex A., Classroom Observation Systems in Preparing School Personnel: An Annotated Bibliography, ERIC Clearinghouse on Teacher Education, Washington, D.C., 1970.

223. Shelton, John L., "The relationship of various levels of accurate empathy to rehabilitation outcome as measured by the interaction scale," Dissertation Abstracts International, 1970(Feb.), Vol. 30(8-B), p3876.

224. Smith, Louis M., Participant Observation and Evaluation Strategies," From Symposium "Parcipant Observation and Curriculum: Research and Evaluation," American Educational Research Association, New York, New York, February, 1971.

225. Stevens, Frederick L., "Client satisfaction as a counseling outcome variable," Dissertation Abstracts International, 1972(Jan.), Vol. 32(7-B), pp4229-4230.

226. Teresa, Joseph G., and Francis, John B., "A Multi-Variate Analysis of Teacher-Student Interpretations of Non-Verbal Cues: The Measurement of Visuo-Gestural Channel Expression," Presented at the American Educational Research Association, Chicago, Illinois, April, 1972.

227. Tom, Alan, and Woodley, Celeste P., Classroom Observation and Analysis: Sources for Social Studies Teachers, ERIC Clearinghouse for Social Studies/Social Science Education Consortium, Inc., Boulder, Colo., 1971.

228. Troth, William A., Hall, Gwendolyn L., and Seals, James M., "Counselor-counseless interaction," Journal of Counseling Psychology, 1971(Jan.), Vol. 18(1), pp77-80.

229. Watson, J. P., "A repertory grid method of studying groups," British Journal of Psychiatry, 1970(Sep.), Vol. 117(538), pp309-318.

230. ..... "A measure of therapist-patient understanding," British Journal of Psychiatry, 1970(Sep.), Vol. 117(538), pp319-321.

231. DeVito, Joseph A., Communication: Concepts and Processes, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1971.

232. Durand de Bousingen, R., and Timmons, G. Dean, "Doctor-patient communication," Acta Symbolica, 1972(Fal.), Vol. 3(2), pp118-122.

233. Galloway, Charles M., An Analysis of Theories and Research in Nonverbal Communication, ERIC Clearinghouse on Teacher Education, Washington, D.C., February, 1972.

234. Giffin, Kim, et al., Eds., Basic Readings in Interperonal Communication, Harper & Row, New York, 1971.

235. Jacobs, Jerry, "Perplexity, confusion and suspicion: A study of selected forms of doctor-patient interaction," Social Science & Medicine, 1971(Apr.), Vol. 5(2), pp151-157.

236. Knapp, Mark L., and Harrison, Randall P., "Observing and Recording Nonverbal Data in Human Transactions," Paper presented at the Annual Convention of the Speech Communication Association, Chicago, 1972.

237. Larson, Carl E., and Dance, Frank E. X., Eds., Perspectives on Communication: Colloquium Proceedings, Winconsin University, Speech Communication Center, Milwaukee, 1968.

238. McCroskey, James C., et al., An Introduction to Interpersonal Communication, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1971.

239. Mortensen, C. David, Communication: The Study of Human Interaction, McGraw-Hill Book Company, New York, New York, 1972.

240. Nolan, Michael J., "Nonverbal Expressive Behavior in an Integrated Model of Human Communication," Paper presented at the Annual Meeting of the International Communication Association, Montreal, April, 1973.

241. Pace, R. Wayne, et al., Communicating Interpersonally: A Reader, Charles E. Merrill Publishing Co., Columbus, Ohio, 1973.

242. Schalock, H. Del, An Overview of the Teaching Research System for the Description of Teacher Behavior in Context, Oregon State System of Higher Education, Teaching Research Division, Monmouth, December, 1967.

243. Smith, Elbert J., Ed., Registry of Communication Research: An Identification of Selected Communication Research Projects in the Academic Community, National Center of Communication Arts and Sciences, Denver, Colorado, 1971.

Audio/Visual Recording Systems

244. Eisler, Richard M., Hersen, Michel and Agras, W. Stewart, "A method for the controlled observation of nonverbal interpersonal behavior," Behavior Therapy, 1973(May), Vol. 4(3), pp420-425.

245. Horner, Walter R., et al., The Development Test, and Evaluation of Three Pilot Performance Reference Scales, American Institutes for Research in the Behavioral Sciences, Pittsburgh, Pa., Aug., 1970.

246. Knoop, Patricia A., and Welde, William L., Automated pilot performance assessment in the T-37: A feasibility study, Wright-Patterson Air Force Base, O, Human Resources Lab., No. AFHRL-TR-72-6, 1973.

Interviewing

247. Alpert, Murray, "Television tape for evaluation of treatment response," Psychosomatics, 1970(Sep.), Vol. 11(5), pp467-469.

248. Asher, James J., "How the applicant's appearance affects the reliability and validity of the interview," Educational & Psychological Measurement, 1970(Fal), Vol. 30(3), pp687-695.

249. Bloch, Ellin L., Goodstein, Leonard D., Jourard, Sidney M., and Jaffe, Peggy E., "Comment on "Influence of an interviewer's disclosure on the self-disclosing behavior of interviewees," Journal of Counseling Psychology, 1971(Nov.), Vol. 18(6), pp595-600.

250. Delahanty, David, "Three aspects of nonverbal communication in the interview," Personnel Journal, 1970(Sep.), Vol. 49(9), pp757-759.

251. DiBennardo, Frank R., "The effect of problem locus and problem focus upon self-relevant content and rapport in relation to interview sequence: An investigation of dynamics of communication in the counseling interview," Dissertation Abstracts International, 1971(Nov.), Vol. 32(5-B), p2999.

252. Dickes, Robert; Simons, Richard C., and Weisfogel, Jerry, "Difficulties in diagnosis introduced by unconscious factors in the interview," Psychiatric Quarterly, 1970(Jan.), Vol. 44(1), pp55-90.

253. Dumont, Florent R., "Certain verbal characteristics and visual interaction in an interview setting as a function of interpersonal distance, room size and induced stress," Dissertation Abstracts International, 1971(Oct.), Vol. 32(4-B), p2417.

254. Elsenrath, Dennis E., Coker, David L., and Martinson, William D., "Microteaching interviewing skills," Journal of Counseling Psychology, 1972(Mar.), Vol. 19(2), pp150-155.

255. Enelow, Allen J. and Swisher, Scott N., "Interviewing and patient care, Oxford University Press, New York, New York.

256. Goodwin, Dwight L., and Garvey, William P., "A method of training psychologists as behavioral consultants," Journal of Consulting & Clinical Psychology, 1971(Dec.), Vol. 37(3), pp355-363.

257. Helfer, R., and Hess, J., "An experimental model for making objective measurements of interviewing skills," Journal of Clinical Psychology, 1970(Jul.), Vol. 26(3), pp327-331.

258. Kadushin, Alfred, The Social Work Interview, Columbia University Press, New York, 1972.

259. Kaul, Theodore J., and Schmidt, Lyle D., "Dimensions of interviewer trustworthiness," Journal of Counseling Psychology, 1971 (Nov.), Vol. 18(6), pp542-548.

260. Kimmel, Paul R., "Interviewing multi-national respondents," Catalog of Selected Documents in Psychology, 1971(Fal), Vol. 1,27.

261. Linder, Ralph, "A seminar in doctor-patient relationships in a rehabilitation medicine setting," Psychosomatics, 1969(Nov.), Vol. 10(6), pp354-359.

262. Miller, Arthur A., Curstein, Alvin G., and Leider, Robert J., "Teaching and evaluation of diagnostic skills," Archives of General Psychiatry, 1971(Mar.), Vol. 24(3), pp255-259.

263. Packwood, William T., and Parker, Clyde A., "A method for rating counselor social reinforcement and persuasion," Journal of Counseling Psychology, 1973(Jan.), Vol. 20(1), pp38-43.

264. Salzman Carl, Shader, Richard I., Scott, Dorothy A., and Binstock, William, "Interviewer anger and patient dropout in walk-in clinic," Comprehensive Psychiatry, 1970(May), Vol. 11(3), pp.267-273.

265. Singer, Paul R., and Muslin, Hyman L., "Evaluation and teaching of psychiatric interviewing," Comprehensive Psychiatry, 1970(Jul.), Vol. 11(4), pp371-376.

266. Smith, Edward W., "Non-verbal communication in interviewer-interviewee dyads as a function of the participants A-B status," Dissertation Abstracts International, 1970(Feb.), Vol. 30(8-B) pp3877-3878.

267. Strong, Stanley R., Taylor, Ronald G., Bratton, Joseph ., and Loper, Rodney G., "Nonverbal behavior and perceived counselor characteristics," Journal of Counseling Psychology, 1971(Nov.), Vol. 18(6), pp554-561.

268. Thorne, F. C., "Psychological 'twenty questions:' A method for teaching diagnostic interviewing," Journal of Clinical Psychology, 1970(Jul.), Vol. 26(3), pp331-334.

269. Wittmer, Joe, "An objective scale for content-analysis of the counselor's interview behavior," Counselor Education & Supervision, 1971(Spr.), Vol. 10(3), pp283,290.

Performance Testing

270. Boyd, Joseph L., Jr., and Shimberg, Benjamin, Handbook of Performance Testing: A Practical Guide for Test Makers, Educational Testing Service, Princeton, New Jersey, January, 1971.

271. Engel, John D., An approach to standardizing human performance assessment, HumRRO Professional Paper, 1970(Oct.), No. 26-70.

272. McLaurin, William A., et al., Rational for and Development of a Battery of Performance Tests for Vocational Selection

273. Osborn, William C., Developing performance tests for training evaluation, HumRRO Technical Report 1973(Feb.), No. PP3-73.

274. Quirk, Thomas J., "Performance Tests for Beginning Teachers: Why All the Fuss?" Educational Technology, May 1973, Vol. 13, pp14-16.

275. Ramsey, H. Rudy, "The semiautomated test system: A tool for standardized performance testing," USN AMI Army-Navy Joint Report, 1969(Nov.), No. 1092.

276. Vineberg, Robert and Taylor, Elaine N., The Interchangeability of Job Sample Tests and Job Knowledge Tests in Four Army Jobs, HumRRO, Monterey, California

277. Wilson, Clark L., et al., A Manual for Use in the Preparation and Administration of Practical Performance Tests, Office of Naval Research, Washington, D.C., 1971.

Simulation

278. Barton, Richard F., "Designing simulation laboratories," Simulation & Games, 1972(Jun.), Vol. 3(2), pp219-225.

279. Chesler, David J., Computer-Assisted Performance Evaluation for Navy Anti-Air Warfare Training: Concepts, Methods and Constraints, Naval Personnel Research Activity, San Diego, California, May, 1971.

280. .....Application and Utilization of Training Aids and Devices: Simulated Exercises and Trainee Performance Evaluation. Final Report, Naval Personnel and Training Research Lab., San Diego, California, September, 1972.

281. Cruickshank, Donald R., and Broadbent, Frank W., Simulation in Preparing School Personnel, ERIC Clearinghouse on Teacher Education, Wahsington, D.C., February, 1970.

282. Fredrickson, Ronald H., and Popken, Charles F., Simulated Administration of a Regular Guidance Operation (SARGO), Massachusetts University, Amherst, School of Education, August, 1970.

283. Grimsley, Douglas L., Acquisition, Retention, and Retraining: Effects of High and Low Fidelity in Training Devices. Technical Report 69-1, George Washington University, Human Resources Research Office, Alexandria, Virginia, February, 1969.

284. Grunwald, Walter, An Investigation of the Effectiveness of Training Devices With Varying Degrees of Fidelity, Oklahoma Univ., Norman, 1968.

285. O'Leary, Lawrence R., "Fair employment, sound psychometric practice, and reality: A dilemma and a partial solution," American Psychologist, 1973(Feb.), Vol. 28(2), pp147-150.

286. Osborn, William C., An approach to the development of synthetic performance tests for use in training evaluation, HumRRO Professional Paper, 1970(Dec.), No. 30-70.

287. Pieper, William J., et al., Learner-Centered Instruction (LCI). Volume 5. Description of the Job Performance Test, Applied Science Associates, Inc., Valencia, Pennylsvania, June, 1969.

288. Reibling, Louis Albert, The Phantom Strikes: The Use of Simulation to Teach the Radiologic Technology Student to Radiograph the Gall Bladder, St. Louis University, 1972.

Gaming

289. Jankowicz, Andres Z., "Feedback for Learning in Business Games," Simulation & Games, Vol. 4(2), June, 1973.

290. Kersch, Susan F., Personnel Services Review. Series 1: Innovations in the Training and Supervision of Counselors. Simulation Gaming, ERIC Clearinghouse on Counseling and Personnel Services, Ann Arbor, Michigan, March, 1970.

291. Kidder, Steven J., and Nafziger, Alyce W. (Eds.), "Proceedings of the National Gaming Council's eleventh annual symposium," Center for Social Organization of Schools Report, Johns Hopkins Univ., 1972 (Nov.), No. 143.

292. Kruper, Karen R., Communication Games: Participant's Manual, The Free Press, New York, New York, 1973.

293. Parry, Scott B., "The name of the game . . . is simulation," Training & Development Journal, 1971(Feb.), Vol. 25(2), pp28-32.

294. Stadsklev, Ronald, A System for Analyzing Social Simulations and Educational Games (SAS) or Games Analysis System (GAS), Concordia Teachers College, Seward, Nebraska, 1969.

295. Varenhorst, Barbara B., "Game theory, simulations and group counseling," Educational Technology, 1973(Feb.,), Vol. 13(2), pp40-43.

Role-Playing

296. Authony, William A., and Wain, Harold J., "An investigation of the outcome of empathy training for medical corpsmen," Psychological Aspects of Disability, 1971(Jul.), Vol. 18(2), pp86-88.

297. Clore, Gerald L., Jeffery, Katherine McMillan, "Emotional Role Playing, Attitude Change, and Attraction Toward a Disabled Person," Paper presented at the Midwestern Psychological Association Convention in Detroit, Michigan, May, 1971.

298. Corsini, Raymond J., et al., Roleplaying in Business and Industry, The Free Press of Glencoe, Inc., 1961.

299. Foegen, J. H., "Super role-playing: Labor and management," Personnel Journal, 1971(Aug.), Vol. 50(8), pp611-617.

300. Froehlich, Dean K., An Experimental Criterion of Cross-Cultural Interaction Effectiveness: A Study of Military Advisors and Counterparts, George Washington University, HumRRO Office, Alexandria, Virginia, December, 1968.

301. Gunter, Neil, "Self-role perception and cross-role perception of school counselors, school social workers, and school psychologists," Dissertation Abstracts International, 1970(Jun.), Vol. 30 (12-A), pp5235-5236.

302. Harris, Philip R., and Harris, Dorothy L., "Training for cultural understanding," Training & Development Journal, 1972(May), Vol. 26(5), pp8-10.

303. Kraemer, Alfred J., A Cultural Self-Awareness Approach to Improving Intercultural Communication Skills, HumRRO, Alexandria, Virginia, April, 1973.

304. MacLennan, Beryce W., "Simulated situations in group psychoterapy training," International Journal of Group Psychotherapy, 1971(Jul.), Vol. 21(3), pp330-332.

305. Moment, David, et al., Role Development and Interpersonal Competence, Harvard University, 1963.

306. Shaftel, Fannie R., et al., Role-playing for social values: decision-making in the social studies, Prentice-Hall, 1967.

307. Stewart, Edward C., et al., Simulating Intercultural Communications through Role-Playing, HumRRO Technical Report 69-7, May, 1969.

308. Wolfe, Lawrence A., "A comparison of the effectiveness of teaching basic interviewing skills by three group methods: Modeling with role-playing,role playing, and discussion," Dissertation Abstracts International, 1970(Sep.), Vol. 31(3-A), pp1029-1030.

In-Basket Exercise

309. Cook, Edward M., "The In-Basket Exercise: For Secretarial Training?" Training and Development Journal, July, 1973.

Natural Setting Techniques

310. Campion, James E., "Work sampling for personnel selection," Journal of Applied Psychology, 1972(Feb.), Vol. 56(1), pp40-44.

311. Pinsky, Paul D.,"A Mathematical Model of Testing for Instructional Management Purposes," From symposium "Designing Instructional Systems with Longitudinal Testing Using Item Sample Techniques," Annual meeting, American Educational Research Association, Minn., Minnesota, March, 1970.

312. Popham, W. James, Performance Tests of Instructor Competence for Trade and Technical Education, Final Report., Office of Education, Washington, D.C, June, 1968.

313. Romberg, Thomas A., Achievement Monitoring Via Item Sampling (Revised), Wisconsin Univ., Research and Development Center for Cognitive Learning, Madison, March, 1970.

314. Siegel, Arthur I., and Bergman, Brian A., Nonverbal and Culture Fair Fair Performance Prediction Procedures: I. Background, Test Development, and Initial Results, Applied Psychological Services, Inc., Wayne, Pennsylvania, June, 1972.

Adaptive Performance Testing

315. Ferguson, Richard L., Computer-Assisted Criterion-Referenced Measurement, Pittsburgh University, Learning Research and Development Center, Pennsylvania, 1969.

316. Johnson, Edward S., Can a Computer Adequately Administer a Complex Problem Solving Task to Individual Human Subjects?, Yale Univ., Department of Psychology Technical Report 12, June,1966.

317. Lord, Frederic M., "A Theoretical Study of the Measurement Effectiveness of Flexilevel Tests," Educational & Psychological Measurement, 1971, Vol. 31, pp805-813.

318. ..... The Self-Scoring Flexilevel Test, Educational Testing Service, Princeton, New Jersey, July, 1970.

319. Roid, Gale H., "Branching methods for constructing psychological test scales," Dissertation Abstracts International, 1970(Sep.), Vol. 31(3-B), p1571.

320. Wargo, Michael J., "Time-estimation skill assessment: A demonstration of adaptive assessment techniques," Human Factors, 1971 (Apr.), Vol. 13(2), pp163-172.

321. Waters, Carrie W., Comparison of Computer-Simulated Conventional and Branching Tests, Behavior and Systems Research Lab. (Army) Arlington, Virginia, March, 1970.

Measurement Category IV: BEHAVIORAL RECORD AUDIT

Unobtrusive Measures

322. Sechrest, Lee. "Unobtrusive Measures in Data Collection." Proceedings of the Academy of Management, 1971 (Aug.), 58-66.

323. Behring, Daniel W. "Adaptive Functioning: A Rationale for the Prediction of Achievement in Nursing Education." Dissertation Abstracts International, 1970 (Sep.), Vol. 31 (3-A). 1065.

324. Bledsoe, Joseph C. "The Prediction of Teacher Competence: A Comparison of two Multivariate Statistical Techniques." Multivariate Behavioral Research, 1973 (Jan.), Vol. 8(1), 3-22.

325. Boyles, Wiley R., Prunkl, Peter R., & Wahlberg, James L. Combat Aviator Criterion Development. HumRRO Professional Paper, 1969 (Nov.), No. 34-69, 10 p.

326. Cox, Steven G. "Do Educational Measures Predict Vocational Success?" Vocational Guidance Quarterly, 1971 (Jun.), Vol. 19(4), 271-274.

327. Gavin, James F. "Ability, Effort and Role Perception as Antecedents of Job Performance." Dissertation Abstracts International, 1970 (Feb.), Vol. 30(8-B), 3901-3902.

328. Goebel, Ronald A., Baum, David R., & Hagin, William V. Using a Ground Trainer in a job Sample Approach to Predicting Pilot Performance. USAF AFHRL Technical Report 71-50, Nov 1971.

329. Goldman, Leo. "Tests and Counseling: The Marriage that Failed." Measurement & Evaluation in Guidance, 1972 (Jan.), Vol. 4(4), 213-220.

330. Korman, Abraham K. "The Prediction of Managerial Performance: A Preview." Studies in Personnel Psychology, 1970 (Apr.), Vol. 2(1), 4-26.

331. Lau, Alan W., et al. A Non-Cognitive Test Battery as a Predictor of Class "A" School Performance. Naval Personnel Research Activity, San Diego, Calif. March 1970.

332. McMorris, Robert F. & Ambrosino, Robert J. "Self-Report Predictors: A Reminder." Journal of Educational Measurement, 1973 (Spr.), Vol. 10(1), 13-17.

333. Pike, Lewis W., et al. Prediction of Job Performance for Negro and White Medical Technicians; Development of the Instrumentation. Civil Service Commission, Washington D.C., PR-69-4, April 1969.

334. Price, Phillip B. et al, Measurement and Predictors of Physician Performance: Two Decades of Intermittently Sustained Research. National Institutes of Health, Bethesda, Md. 1971, 169 p.

335. Resnikoff, Arthur. "A Measure of Discrimination for the Prediction of Counselor Effects." Journal of Counseling Psychology, 1972 (Sep.), Vol. 19(5). 464-465.

336. Sabet-Sharghi, Zabiholah. "The Relationship Between Motivational Orientation, Role Perception, and Job Performance." Dissertation Abstracts International, 1971 (Sep.), Vol. 32(3-B), 1893-1894.

337. Weinstein, Alan G. "Predicting Behavior From Attitudes". Public Opinion Quarterly, 1972 (Fall), Vol. 36(3), 355-360.

338. Agras, W. Stewart. "Toward the Certification of Behavior Therapists?" Journal of Applied Behavior Analysis, 1973 (Spring), Vol. 6(1), 167-173.

339. "Certification in Allied Health Professions. 1971 Conference Proceedings." Proceedings of the Invitational Conference on Certification in Allied Health Professions held at University of Maryland Continuing Education Center, Sept. 7-10, 1971.

340. Miller, George E., et al. The Orthopaedic Training Study. Final Report. Illinois University, Chicago. Center for the Study of Medical Education. 1968.

341. Report of a National Study of Cytotechnologists: Education and Performance Relationships. National Council on Medical Technology Education, Memphis, Tenn. June 1968.

342. Morgenstern, Alan L. "A Criticism of Psychiatry's Board Examinations." American Journal of Psychiatry, 1970 (July), Vol. 127(1), 33-42.

343. Taylor, Robert L. & Torrey E. Fuller. "The Pseudo-Regulation of American Psychiatry." American Journal of Psychiatry, 1972 (Dec.), Vol. 129(6), 658-663.

344. "Evaluation of Graduate Physician Manpower Education". Association for Hospital Medical Education, Wash., D.C. Edited transcript of Seminar (Chicago, Ill. Feb 5, 1969), 5 February 1969. 146 p.

345. Athelstan, Gary T., et al, Pilot Study of the Practice of Physical Medicine and Rehabilitation. Final Report. American Rehabilitation Foundation, Minneapolis, Minn. March 1970.

346. Cline, David W. & Garrard, Judith N. "A Medical Interviewing Course: Objectives, Techniques, and Assessment." American Journal of Psychiatry, 1973 (May), Vol. 130(5). 574-578.

347. Elstein, Arthur S., & Shulman, Lee S. "A Method for the Study of Medical Thinking and Problem Solving." Michigan State University, East Lansing, February 1971. Paper presented at the Annual Meeting of the American Educational Research Assoc., N.Y. Feb 1971.

348. Haley, John V. Effects of Training on the Test of Diagnostic Skills, Publication No. 30, Loyola University, Chicago, Ill. Psychometric Lab., 1963.

349. Jason, Hilliard, et al. "New Approaches to Teaching Basic Interview Skills to Medical Students." American Journal of Psychiatry, 1971 (April), Vol. 127(10), 1404-1407.

350. Levine, Harold G. & Noak, John R. The Evaluation of Complex Educational Outcomes. Illinois State Office of the Superintendent of Public Instruction, Springfield Dept. of Educational Research; Illinois Univ., Chicago. College of Medicine, 1968.

351. Meyer, Thomas. A Feasibility Study in Determining Individual Practice Profiles of Physicians as a Basis for Continuing Education of These Physicians Utilizing a Postgraduate Perceptor Technique. Final Report. Wisconsin University, Madison. 1970.

352. Price, Phillip B. Measurement and Predictors of Physician Performance: Two Decades of Intermittently Sustained Research. Salt Lake City: Aaron Press, 1971.

353. Rimoldi, H.J.A. Evaluation and Training of Clinical Diagnostic Skills, Publication No. 41. Loyola Univ., Chicago, Ill., Psychometric Lab. 1963.

354. Taylor, C.W., et al. Synthesis of Multiple Criteria of Physician Performance. Utah Univ., Salt Lake City., Dept. of Psychology, November 1968.

355. Thurnblad, Robert J., Muslin, Hyman & Loesch, John, "A Test of Clinical Learning by Medical Students." American Journal of Psychiatry, 1973 (May), Vol. 130(5), 568-570.

356. Waldron, John. "Teaching Communication Skills in Medical School." American Journal of Psychiatry, 1973 (May), Vol. 130(5), 579-581.

357. Alvares, Kenneth M. & Hulin, Charles L. "Two Explanations of Temporal Changes in Ability-Skill Relationships: A Literature Review and Theoretical Analysis." Human Factors, 1972 (August), Vol. 14(4), 295-308.

358. Andrulis, Richard S. "Construct Validation of a Standardized Achievement Test." Proceedings of the Annual Convention of the American Psychological Association, 1972, Vol. 7 (Pt. 1), 5-6.

359. Bergman, Brian A. & Siegel, Arthur I. Training Evaluation and Student Achievement Measurement: A Review of the Literature. USAF AFHRL Technical Report, No. 72-3, 1972 (January), 1-57.

360. Bond, Nicholas A. & Rigney, Joseph W. Measurement of Training Outcomes. University of Southern California, Department of Psychology Technical Report, No. 66, 1970 (June), 34 p.

361. Gronlund, Norman E. Measurement and Evaluation in Teaching. (2nd Ed.) New York, Macmillan, 1971, ix, 545 p.

362. Siegel, Arthur I., et al. Some Techniques for the Evaluation of Technical Training Courses and Students. USAF AFHRL Technical Report, No. 72-15, 1972 (February), 137 p.

363. Bergman, Brian A., & Siegel, Arthur I. Training Evaluation and Student Achievement Measurement: A Review of the Literature. AFHRL-TR-72-3, January 1972, 67p.

364. Bock, R. Darrell & Wood, Robert. "Test Theory." Annual Review of Psychology, 1971, Vol. 22, 193-224.

365. Campbell, John P. "Personnel Training and Development." Annual Review of Psychology, 1971, Vol. 22, 565-602.

366. Cliff, Norman. "Scaling." Annual Review of Psychology, 1973, Vol. 24, 473-506.

367. Hieronymus, A.N. "Today's Testing: What do we Know How to do?" Proceedings of the 1971 Invitational Conference on Testing Problems, 1972, 57-68.

368. Holmen, Milton G. & Docter, Richard F. Educational and Psychological Testing: A Study of the Industry and its Practices. New York, Russell Sage Foundation, 1972, ix, 218 p.

369. Hutt, S.J., & Hutt, Corinne. Direct Observation and Measurement of Behavior. Springfield, Ill.: Charles C. Thomas, 1970, xii, 224 p.

370. Leigh, Terry. A Selected and Annotated Bibliography on Evaluating Performance of College Faculty Members. Kentucky University, Lexington, Report No. Occas-pap-8, 1969, 30 p.

371. Mather, William G., et al. "Man, His Job, and the Environment: A Review and Annotated Bibliography of Selected Recent Research on Human Performance." Catalog of Selected Documents in Psychology, 1971, (Fall), Vol. 1, 28-29.

372. Schalock, H. Appendix Z. Classes of Measures Used in the Behavioral Sciences, the Nature of the Data that Derive from Them, and Some Comments as to the Advantages and Disadvantages of Each. Northwest Regional Euucational Lab, Portland, Ore., October 1968. Appendix Z in a Competency Based, Field Centered, Systems Approach to Elementary Teacher Education: Final Report, Vol. III, P. 244-249.

373. Sjogren, Douglas D. "Measurement Techniques in Evaluation." Review of Educational Research, 1970, (April), Vol. 40(2), 301-320.

374. Tiedeman, Herman R. Fundamentals of Psychological and Educational Measurement. Springfield, Ill.: Charles C. Thomas, 1972, ix, 134 p.

375. Thorndike, R.L. (Ed.) Educational Measurement, (2nd ed.), Washington, D.C.: American Council on Education, 19.1.

376. Abrahams, Norman M., & Alf, Edward. "Pratfalls in Moderator Research." Journal of Applied Psychology, 1972 (June), Vol. 56(3), 245-251.

377. Cirino-Gerena, Gabriel. "Toward a Model for Test Prediction." American Psychologist, 1972, (August), Vol. 27(8), 759-760.

378. Cunningham, J.W., "Ergometrics": A Systematic Approach to Some Educational Problems. North Carolina State Univ., Raleigh, Center for Occupational Education, Report No. Cen-Mong-7, 1971.

379. Cunningham, J.W., et al., The Development of the Occupation Analysis Inventory: An "Ergometric" Approach to an Educational Problem. North Carolina State Univ., Raleigh, Center for Occupational Education, Report No. Cen-Res-Monog-6, 1971.

380. Constructing and Using Achievement Tests: A Guide for Navy Instructors. Department of the Navy, Washington, D.C., Bureau of Naval Personnel, NAVPERS-16808-B, 1971.

381. Dirks, H. "The Methodological Basis of Work, Performance and Personality Assessment Systems." Psychologia Africana, 1970 (October), Vol. 13(2-3), 124-137.

382. du Mas, Frank M. "Superpsychometrics: Ultrareliability and Hypervalidity." Journal of Psychology, 1972, (July), Vol. 81(2), 329-342.

383. "Invitational Conference on Testing Problems." Educational Testing Service, Princeton, N.J., 28 October 1967. Proceedings of the 1967 Invitational Conference on Testing Problems, N.Y. October 1967.

384. Ekman, Gosta. "Comparative Studies on Multidimensional Scaling and Related Techniques." Reports from the Psychological Laboratories, U. Stockholm, 1970, (November), Suppl. 3, 28 p.

385. Fox, A.M. & Brookshire, William K., "Defining Effective College Teaching Using the Delphi Technique and Multiple Linear Regression." Paper presented at the 1971 Convention of the American Educational Research Association, New York City.

386. Gardner, Eric F. Measurement in Education: Interpreting Achievement Profiles - Uses and Warnings. National Council on Measurement in Education, East Lansing, Mich., January 1970.

387. Ghiselli, Edwin E., "Comment on the Use of Moderator Variables." Journal of Applied Psychology, 1972, (June), Vol. 56(3), 270.

388. Kaiser, Henry F. & Carter, Harold D. "A Geometric Representation of the Notions of Reliability, Relevance, and Validity." California Journal of Educational Research, 1971, (May), Vol. 22(3), 122-125.

389. Korman, Abraham K., et al, "The Generality of the Characteristics of Competent People." Journal of Vocational Behavior, 1971, Vol. 1, 201-208.

390. Livingston, Samuel A. "Criterion-Referenced Applications of Classical Test Theory." Journal of Educational Measurement, 1972, (Spring), Vol. 9(1), 13-26.

391. Lord, Frederic M. "A Theoretical Study of the Measurement Effectiveness of Flexilevel Tests." Educational & Psychological Measurement, 1971 (Winter), Vol. 31(4), 805-813.

392. Mannen, Howard A, (Ed.) "Current Research Techniques in Military Personnel Assessment." Proceedings of Annual Conference, Military Testing Association (12th, Sep 14-18, 1970).

393. Marcus, Edward E., "What do you Mean, "Evaluation?" Personnel Journal, 1971 (May), Vol. 50(5), 354-358.

394. McClelland, David C. "Testing for Competence Rather Than for "Intelligence." American Psychologist, Vol. 28, No. 1, Jan 1973.

395. Meese, M. Kathryn, "A Model for Assessing Complex Educational Outcomes." Paper presented at the Annual Meeting of the American Educational Research Association, New York, February 1971.

396. Millman, Jason, "Psychometric Characteristics of Performance Tests of Teaching Effectiveness." Paper presented at annual meeting of the American Educational Research Association, New Orleans, La., February 25 1973.

397. Shavelston, Richard J, et al. "Criterion-Referenced Testing: Comments on Reliability." Journal of Educational Measurement, 1972 (Summer), Vol. 9(2), 133-140.

398. Theologus, George C. & Fleishman, Edwin A. "Development of a Taxonomy of Human Performance: Validation Study of Ability Scales for Classifying Human Tasks." Catalog of Selected Documents in Psychology, 1973 (Winter), No. 3,29.

399. Theologus, George C., et al. "Development of a Taxonomy of Human Performance: A Feasibility Study of Ability Dimensions for Classifying Human Tasks." Catalog of Selected Documents in Psychology, 1973, (Winter), No. 3,25-26.

400. Tucker, Ledyard R. "An Inter-Battery Factor Analytic Approach to Linear Predictive Systems." Educational & Psychological Measurement, 1973, (Spring), Vol. 33(1), 77-97.

401. Uhlaner, J.E., "Human Performance Effectiveness and the Systems Measurement Bed." Journal of Applied Psychology, 1972, (June), Vol. 56(3), 202-210.

402. Velicer, Wayne F. "A Comment on the General Inapplicability of Ghiselli's Moderator System for Two Predictors." Journal of Applied Psychology, 1972, (June), Vol. 56(3), 262-265.

403. Velicer, Wayne F. "The Moderator Variable Viewed as Heterogeneous Regression." Journal of Applied Psychology, 1972, (June), Vol. 56(3), 266-269.

404. Wentling, Tim L. "Measuring the Achievement of Competencies." Educational Technology, May 1973, Vol. 13, 48-50.

TEST DIRECTORIES

405. Boyd, Joseph L, Jr., & Shimberg, Benjamin. Directory of Achievement Tests for Occupational Education. Educational Testing Service, Princeton, N.J., April 1971.

406. Hoepfner, Ralph, Ed, et al. CSE-RBS Test Evaluations: Tests of Higher-Order Cognitive, Affective, and Interpersonal Skills. California Univ., Los Angeles. Center for the Study of Evaluation, Research for Better Schools, Inc., Philadelphia, Pa. 1972.

407. Index and Description of Tests. New York Univ., N.Y. Institute for Developmental Studies, May 1965.